DESCRIÇÃO SEMÂNTICA DE OBJETOS EM IMAGENS BASEADA NA TEORIA DOS PROTÓTIPOS

OMAR VIDAL PINO

DESCRIÇÃO SEMÂNTICA DE OBJETOS EM IMAGENS BASEADA NA TEORIA DOS PROTÓTIPOS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Orientador: Mario Fernando Montenegro Campos
Coorientador: Erickson Rangel do Nascimento

Belo Horizonte
Fevereiro de 2020

Vidal Pino, Omar.

V648d Descrição semântica de objetos em imagens baseada na Teoria dos Protótipos [manuscrito] / Omar Vidal Pino.- 2020.
xxxiv, 235 f. il.

Orientador: Mário Fernando Montenegro Campos.
Coorientador: Erickson Rangel do Nascimento.
Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.159-178
.
1. Computação – Teses. 2. Teoria dos protótipos – Teses. 3. Aprendizado profundo – Teses. 4. Visão computacional – Teses. I. Campos, Mário Fernando Montenegro. II. Nascimento, Erickson Rangel do .III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV.Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Descrição semântica de objetos em imagens baseada na Teoria dos Protótipos

OMAR VIDAL PINO

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MARIO FERNANDO MONTENEGRO CAMPOS - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ERICKSON RANGEL DO NASCIMENTO - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. ANDERSON DE REZENDE ROCHA
Instituto de Computação - UNICAMP

PROF. WAGNER MEIRA JUNIOR
Departamento de Ciência da Computação - UFMG

DR. RENATO JOSÉ MARTINS
Pós-Doutorado - INRIA

PROF. LUIZ CHAIMOWICZ
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 10 de Fevereiro de 2020.

A la memoria de mi amado padre...

Agradecimentos

Fazer um doutorado exige muito esforço e sacrifício. Fazê-lo fora do seu país, longe dos seus parentes e da zona de conforto, com certeza, torná-lo ainda mais difícil. Mas, quando tudo acaba, você se sente imensamente feliz e ao mesmo tempo grato por tudo e por todos. Estou eternamente grato por todas aquelas pessoas que me ajudaram, direta e indiretamente, na minha estadia no brasil e no desenvolvimento e conclusão deste trabalho.

Primeiro, gostaria de agradecer a meu orientador Mario Fernando Montenegro Campos por ter apostado em mim e me dar a oportunidade de fazer parte desse excelente ambiente de pesquisa que é o Verlab. Obrigado professor por sua paciência e confiança, mesmo quando durante o processo, as vezes tudo parecia perdido. Também quero agradecer a meu co-orientador Erickson Rangel do Nascimento. Gostaria de registrar minha admiração e sincero agradecimento por todas as lições, orientações e conhecimentos que me foram transmitidos. Foi uma ótima experiência de auto-aperfeiçoamento, crescimento e aprendizado profissional. Obrigado.

Agradeço eternamente à minha família; a meus pais Omar e Elena pelos sacrifícios feitos para que eu pudesse seguir meus sonhos. Minha irmã Lilena por simular minha presença e por ter sido o apoio e a força da família nos momentos mais difíceis e tristes. Por último, mas não menos importante, minha esposa, suporte inesgotável e motor emocional insubstituível; para ti meu amor não tenho uma expressão que realmente mostre minha eterna gratidão, simplesmente: *Gracias*. Para todos vocês, meus sinceros agradecimentos e amor.

Também gostaria agradecer a essa incrível comunidade acadêmica que é a equipe do VerLab; um conjunto de pessoas que fornecem um apoio indispensável, colaborando com suas habilidades e boa vontade. Especificamente, gostaria de expressar minha gratidão para todos aqueles que com suas sugestões, recomendações e críticas me ajudaram no meu trabalho de pesquisa: David Saldaña, Paulo Drews, Levi Vasconcelos, Daniel Balbino, Elerson Rubens, Anderson Rocha, Jhielson Pimentel, Rafael Colares, Ramon Melo, Thiago Gomes e outros que estou esquecendo, a todos minha gratidão.

Aos professores Dr. Eucidio Pimenta Arruda e Dr. Wagner José Corradi Barbosa por me permitirem fazer parte da equipe de trabalho responsável pela construção do primeiro Repositório Institucional da UFMG. Obrigado pela bonita experiência de trabalho e por me dar a oportunidade de retribuir meus agradecimentos à UFMG.

À grande comunidade latina e cubana pelo apoio e diversão que me forneceram durante todo o processo. Especificamente, para aqueles novos amigos que me apoiaram quando mais o precisava e que ao mesmo tempo me deram as lembranças mais belas do Brasil: David Saldaña, Levi Vasconcelos, Rogerio Fonteles, Samuel Sérvulo, Vladimir Portela, Paulo Drews, Igor Campos, Jhielson Pimentel, Clayson Celes, Elisa Ramíres, Michel López, Angel Gutierrez, e Reynier Rojas.

Pelo apoio financeiro e de infraestrutura, meus sinceros agradecimentos a CAPES, CNPq, Fapemig, UFMG, DCC e VeRLab.

"If we want machines to think,
we need to teach them to see."
(Fei-Fei Li)

Resumo

Esta pesquisa tem como objetivo propor um modelo para a descrição semântica das características de objetos a partir de imagens. Apresenta-se uma nova abordagem de descrição semântica de objetos fundamentada na Teoria dos Protótipos. Propõe-se o Modelo Computacional do Protótipo (CPM) para *codificar* e *armazenar* o significado semântico central (protótipo semântico) das categorias de objetos. O modelo CPM é utilizado para representar e construir os protótipos semânticos das categorias de objetos usando as Redes Neuronais Convolucionais (CNN). Propõe-se um *Modelo de Descrição Semântica baseado em Protótipos* que usa o modelo CPM proposto para descrever objetos de maneira a destacar as características que os distinguem dentro de uma categoria. O Descritor Semântico Global proposto (GSDP) constrói assinaturas discriminativas, de baixa dimensionalidade, interpretáveis e que codificam a informação semântica dos objetos por meio dos protótipos semânticos construídos. O descritor semântico GSDP usa a *Camada de Similaridade Prototípica* (PS-Layer) proposta para recuperar o protótipo correspondente à categoria de interesse usando o *princípio de categorização baseado em protótipos*. Os experimentos realizados utilizando conjuntos de dados de domínio público mostraram que: *i)* o modelo CPM proposto simula adequadamente a estrutura interna das categorias; *ii)* a métrica de distância proposta apresenta poder expressivo para capturar a tipicidade do objeto dentro da categoria; *iii)* a classificação semântica baseada em protótipos pode melhorar o desempenho dos modelos CNN de classificação; *iv)* a codificação do descritor semântico proposto é semanticamente interpretável e supera significativamente em desempenho outras codificações globais de imagem em tarefas de agrupamento e classificação.

Palavras-chave: Descrição Semântica, Efeitos Prototípicos, Teoria dos Protótipos, Aprendizagem Profunda, Visão Computacional.

Abstract

This research aims to build a model for semantic description of objects based on features detected in images. We introduce a novel semantic description approach inspired on the Prototype Theory foundations. Inspired by the human approach used for representing categories, we propose a novel *Computational Prototype Model* (CPM) that *encodes* and *stores* the central semantic meaning of the object's category: the semantic prototype. Our CPM model is used to represent and construct the semantic prototypes of object categories using Convolutional Neural Networks (CNN). The proposed *Prototype-based Description Model* uses the CPM model to describe an object highlighting its most distinctive features within the category. Our Global Semantic Descriptor (GSDP) builds discriminative, low-dimensional and semantically interpretable signatures that encode the semantic information of the objects using the constructed semantic prototypes. Our semantic descriptor use the proposed *Prototypical Similarity Layer* (PS-Layer) to *retrieves* the category prototype using the *principle of categorization based on prototypes*. In our experiments, using publicly available datasets, we show that: *i)* the proposed CPM model adequately simulates the internal semantic structure of the categories; *ii)* the proposed semantic distance metric can be understood as the object typicality score within a category; *iii)* our semantic classification method based on prototypes can improve the performance and interpretation of CNN classification models; *iv)* our semantic descriptor encoding significantly outperforms others state-of-the-art image global encoding in clustering and classification tasks.

Keywords: Semantic Description, Prototypicality Effects, Prototype Theory, Deep Learning, Computer Vision.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BRNN Bidirectional Recurrent Neural Network

CIFAR Canadian Institute For Advanced Research

CNN Convolutional Neural Network

CPM Computational Prototype Model

FCN Fully Convolutional Network

FCSS Fully Convolutional Self-Similarity

GCM Generalized Context Model

GSDP Global Semantic Descriptor based on Prototypes

ILSVRC ImageNet Large Scale Visual Recognition Challenge

LIFT Learned Invariant Feature Transform

LSS Local Self-Similarity

LSTM Long Short-Term Memory

MNIST Modified National Institute of Standards and Technology

MPM Multiplicative Prototype Model

NLP Natural Language Processing

NN Neural Network

ReLU Rectified Linear Units

RNN Recurrent Neural Network

SAD Sum of Absolute Difference

SIFT Scale Invariant Feature Transform

UCN Universal Correspondence Network

Lista de Símbolos

θ Um ângulo.

ς Ângulo máximo.

β Ângulo mínimo.

γ Ângulo resultante da bissetriz interna.

v Um vetor.

$\mathbb{R}$ Conjunto dos números reais.

Θ Matriz de direções.

C Conjunto finito de categorias de objetos.

F Conjunto finito de características distintivas de um objeto.

O O universo de objetos.

δ Distância semântica.

$\hat{z}$ Valor semântico.

$\vec{z}$ Representação vetorial do valor semântico.

λ Aplicação não expansiva entre esses espaços métricos da norma l_1.

ρ Aplicação não expansiva entre espaços métricos.

l_1 Soma da Diferença Absoluta.

M_i A média das caraterísticas extraídas para objetos típicos da i-ésima categoria.

μ_{ij} O valor médio da j-ésima caraterística da i-ésima categoria.

ω_{ij} O valor de relevância da j-ésima caraterística da i-ésima categoria.

Ω_i Valores de relevância das características da i-ésima categoria.

P_i Protótipo semântico da i-ésima categoria.

P^O Conjunto de protótipos semânticos do universo de objetos O.

Σ_i O desvio padrão das caraterísticas extraídas dos objetos da i-ésima categoria.

σ_{ij} O desvio padrão da j-ésima característica da i-ésima categoria.

Lista de Equações

Sumário

Capítulo 1

Introdução

A *memória* constitui uma das faculdades mais surpreendentes do ser humano. Geralmente considera-se que a *memória* é a habilidade do cérebro de *codificar*, *armazenar* e *recuperar* a informação (Atkinson & Shiffrin, 1968; Tulving, 2007). Mais especificamente, a *memória semântica* refere-se ao conhecimento do mundo geral que acumulamos ao longo de nossas vidas (McRae & Jones, 2013). Um aspecto relevante da neuroanatomia funcional da memória semântica reside na *representação do significado de um objeto* concreto e as suas propriedades (Martin, 2007). Várias suposições indicam que os seres humanos são capazes de aprender as características mais distintivas de uma categoria ou objeto específico (Thompson-Schill, 2003; Martin, 2007). De fato, os seres humanos começam a formar categorias e abstrações em uma idade muito precoce (Murphy, 2004). A memória semântica envolve a *definição semântica dos objetos* (Tulving, 1992), e consequentemente, o sucesso das tarefas de reconhecimento, classificação e descrição de objetos está causalmente relacionado com o sucesso da tarefa de recuperação do conhecimento aprendido (Tulving, 2007).

Mas, como simular o comportamento da *memória semântica* na representação do conhecimento aprendido das características dos objetos? Como extrair e *codificar* as características relevantes da imagem para *armazenar* a representação do significado (ou representação semântica) de um objeto concreto? Como aprender a *definição semântica dos objetos* para usá-la nas tarefas de reconhecimento, classificação e descrição de objetos? Essas são apenas algumas das questões que ocupam a agenda de trabalho de muitas áreas de investigação. Esta pesquisa –motivada pelo comportamento da *memória semântica*– propõe um modelo que tenta representar a *definição semântica das categoria de objetos*, e propõe como introduzir essa representação semântica na descrição global de objetos no contexto da Visão Computacional e Aprendizagem de Máquina.

Um objetivo constante dessas áreas é desenvolver e aperfeiçoar modelos de aprendizagens com um desempenho que se aproxime, ou supere, ao do ser humano no processamento da informação visual. Os modelos de construção de conhecimento a partir da informação contida na imagem são altamente influenciados pelos métodos usados para a detecção, extração e representação da informação relevante (ou características) da imagem. A extração e a representação de características relevantes da imagem tornou-se um dos objetivos de pesquisa em Visão Computacional por décadas, e constituem tarefas críticas devido ao impacto que geram no desempenho de outras aplicações (por exemplo: reconhecimento de objetos, *stitching* de imagem, classificação de cenas, rastreamento de objetos).

Muitas abordagens vem sendo desenvolvidas para equipar às máquinas com as capacidades de extração e representação (descrição) de boas características (Szeliski, 2010). As características artesanais (Lowe, 2004; Bay et al., 2008; Tola et al., 2008) e as características baseadas na aprendizagem de máquina (Strecha et al., 2012; Perez & Olague, 2013; Simonyan et al., 2014) têm sido utilizadas nas tarefas de descrição de imagens.

O advento das Redes Neurais Convolucionais (*Convolutional Neural Networks (CNNs)*) possibilitou a obtenção de modelos de reconhecimento visual com um comportamento e desempenho semelhantes à *memória semântica* em tarefas de classificação (Simonyan & Zisserman, 2014; Chollet, 2016; He et al., 2016; Szegedy et al., 2016; Howard et al., 2017; Szegedy et al., 2017). Esses resultados desencadearam a tendência do processamento semântico da imagem usando técnicas de aprendizagem profunda em tarefas tais como descrição de imagens em linguagem natural (Karpathy & Fei-Fei, 2015), recuperação da imagem baseada em conteúdo (Zhu et al., 2017) e classificação da cena (Nogueira et al., 2017).

O sucesso dos modelos CNN gerou o surgimento de numerosos descritores CNN. A "família" de Descritores-CNN está constituída por diferentes abordagens que aprendem representações eficazes para descreverem características da imagem (Han et al., 2015; Simo-Serra et al., 2015; Zagoruyko & Komodakis, 2015; Zbontar & LeCun, 2015; Kim et al., 2017). Consequentemente, as representações das características extraídas da imagem usando modelos CNNs de classificação, ou usando descritores CNN, são comumente referidas como *características semânticas* ou *assinaturas semânticas* (Simonyan & Zisserman, 2014; He et al., 2016; Szegedy et al., 2017).

O termo *característica semântica* tem sido amplamente estudado nas Ciências Cognitivas e é definido como a representação dos componentes conceituais básicos do significado de qualquer elemento lexical (Fromkin et al., 2018). No trabalho seminal de Rosch (1975b), a autora analisou a *estrutura semântica* do significado das palavras e

introduziu o conceito de *protótipo semântico* (ou Teoria dos Protótipos). De acordo com Rosch (Rosch, 1975b; Rosch & Mervis, 1975), a *representação do significado semântico da categoria* está relacionada com o *protótipo semântico* da categoria, particularmente para aquelas categorias que designam objetos naturais.

Os modelos CNN de descrição de características (Han et al., 2015; Lin et al., 2016b; Simo-Serra et al., 2015; Zagoruyko & Komodakis, 2015; Zbontar & LeCun, 2015) e os modelos de descrição semântica (Choy et al., 2016; Han et al., 2017; Kim et al., 2017; Rocco et al., 2018) representam a informação semântica das características da imagem usando uma variedade de abordagens diferentes. No entanto, nenhum desses modelos constrói as representações das características codificando a informação da imagem sob a extensa base teórica da Ciência Cognitiva para representar *o significado*.

A imagem pode ser entendida como um artefato que fornece uma representação, mediante uma foto física ou uma foto digital (matriz bidimensional), de uma percepção visual com aparência semelhante aos elementos que compõem a cena; por exemplo um objeto físico, pessoa, rostro, ação, etc. A compreensão semântica da imagem e a extração de conhecimento (ou informação semântica) da imagem mediante um computador está influenciada por como são localizados, reconhecidos e representados esses elementos que a compõem (Guo et al., 2016). Comumente, a interpretação da imagem é realizada a partir das representações das características dos objetos que a compõem, além das representações das relações existentes entre esses objetos (Guo et al., 2016). Ou seja, podemos entender os objetos como os componentes primários de mais alto nível que definem a composição semântica da imagem.

Sob esses supostos, representar *o significado semântico* dos componentes básicos da imagem (os objetos), sugere uma abordagem admissível e um ponto de partida para a compreensão e a representação semântica da informação contida na imagem. Mesmo quando é extenso o fundamento teórico de várias áreas das Ciências Cognitivas como a Neurociência (Thompson-Schill, 2003; Martin, 2007; McRae & Jones, 2013), a Linguística (Rosch, 1975b; Rosch & Mervis, 1975; Geeraerts, 2010) e a Psicologia (Estes, 1986; Minda & Smith, 2002; Zaki et al., 2003) sobre a representação e interpretação da semântica dos objetos pelo ser humano, a elevada abstração de muitos desses fundamentos teóricos dificulta que sejam completamente reproduzidos nas áreas de Aprendizagem de Máquina e Visão Computacional.

Diferente dos modelos existentes na literatura para a representação (descrição) semântica das características da imagem, o presente trabalho propõe introduzir os fundamentos teóricos da semântica cognitiva relacionados com a Teoria dos Protótipos para representar o *significado semântico* da informação contida na imagem.

Teoria dos Protótipos

A Teoria dos Protótipos (Rosch, 1975b; Rosch & Mervis, 1975; Geeraerts, 2010) propõe que os seres humanos definem as categorias em termos de *protótipos abstratos*, definidos como os casos centrais claros (representativos) das categorias (Rosch, 1975b; Rosch & Mervis, 1975). Rosch (Rosch & Mervis, 1975; Rosch, 1975b) mostrou que os humanos aprendem o *significado semântico central da categoria* (ou protótipo da categoria) e o incluem nos processos cognitivos. Em outras palavras, os seres humanos aprendem primeiro o *significado central* da categoria (protótipo semântico da categoria) e posteriormente o *significado periférico* da categoria.

A Teoria dos Protótipos propõe um modelo de categorização e representação semântica alternativo ao modelo tradicional baseado na lógica de Aristóteles (Barnes et al., 1995). Ou seja, contrário à crença de que as categorias são *homogêneas* (onde os elementos e as características possuem a mesma relevância na categoria) e *discretas* (que existe um conjunto numerável e bem definido de categorias e características do objeto), a Teoria dos Protótipos propõe uma concepção das categorias como *heterogêneas* (nas quais existem alguns membros ou características mais representativos que outros dentro da categoria) e *não discretas*. Rosch (1978) considera que não existem definições necessárias e suficientes para definir uma categoria específica (Geeraerts, 2010), e considera à categoria como elementos diferentes que têm um estado desigual de representação semântica (ou tipicidade). Por exemplo, sob essa perspectiva, pode-se entender que o pardal é mais prototípico da categoria ave do que o pinguim.

Rosch (1975b) obteve evidência de que os seres humanos armazenam o significado semântico da categoria de acordo com o grau de representatividade (tipicidade) dos membros da categoria. O *protótipo* foi formalmente definido como os casos centrais claros da categoria (Rosch, 1975b; Geeraerts, 2010). Os atributos desses membros focais incluem, estruturalmente, as propriedades mais salientes que definem à categoria e, inversamente, um membro particular ocupa a posição focal da categoria porque exibe as características mais salientes que caracterizam a categoria.

As Figuras 1.1a e 1.1b apresentam os conceitos principais propostos pela Teoria dos Protótipos para: *i)* a aprendizagem e representação semântica das categorias de objetos, e *ii)* a categorização de objetos baseado em protótipos, respectivamente. O sistema visual humano tem a capacidade de observar elementos de uma mesma categoria e, baseado nas características aprendidas, construir e armazenar uma entidade cognitiva abstrata (protótipo da categoria) que representa o significado semântico central da categoria. Por exemplo, na Figura 1.1a de observar 1 o ser humano constrói 3 (Oliva, 2016).

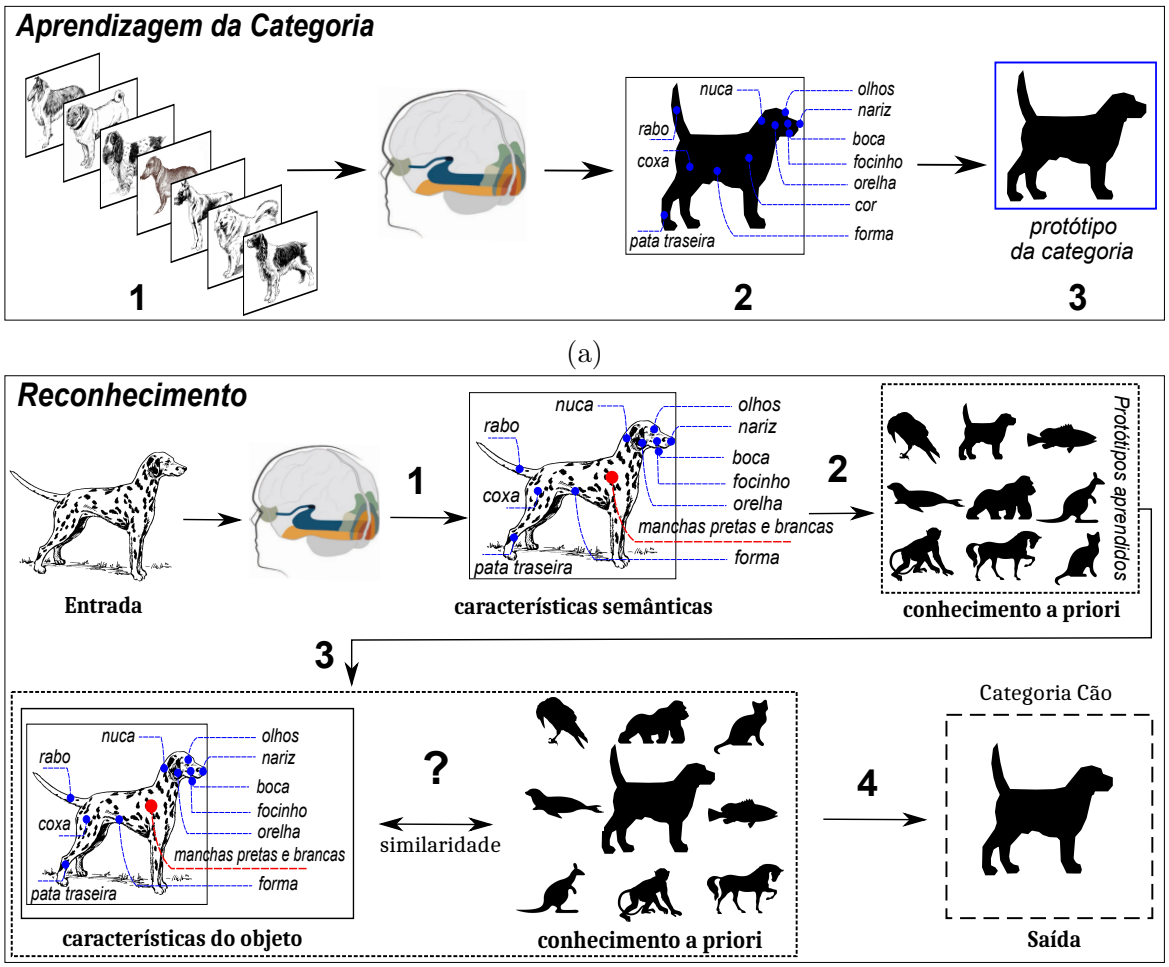


Figura 1.1: *Conceitos principais da Teoria dos Protótipos.* a) *Aprendizagem e representação semântica de uma categoria de objeto*: *1)* exemplares da categoria cão; *2)* características semânticas mais relevantes aprendidas para a categoria cão; *3)* armazenamento apenas do protótipo semântico da categoria. b) *Reconhecimento e categorização baseada em protótipos*: *1)* extração de características; *2)* comparação das características do objeto com os protótipos aprendidos; *3)* reconhecimento do objeto baseado na similaridade com os protótipos das categorias aprendidas; *4)* categorização baseada em protótipos. Fonte: Elaborado pelo autor.

Aliás, as características semânticas aprendidas das categorias de objetos são usadas pelo cérebro humano para *identificar*, *classificar* e *descrever* um objeto concreto (Tulving, 2007). O *conceito de categorização baseado em protótipos* (Rosch & Mervis, 1975; Rosch, 1975b) e o *modelo do protótipo* (Posner & Keele, 1968; Reed, 1972; Homa & Vosburgh, 1976; Zaki et al., 2003) propõem que a execução bem-sucedida da tarefa de reconhecimento de objetos no cérebro humano está intrinsecamente relacionada aos protótipos das categorias aprendidas. De acordo com esse modelo, o homem classifica os objetos com base em sua similaridade com algum protótipo das categorias aprendidas. Na Figura 1.1b dado um objeto de entrada, a categorização baseada em protótipos é realizada seguindo o fluxo de processos 1-4.

Motivação

A Teoria dos Protótipos propõe um modelo de *representação do significado semântico dos objetos* onde a representação do *significado semântico central da categoria* (o protótipo) possui um papel protagonista. Sob essas premissas, a ideia central desta pesquisa consiste em usar os *protótipos* das categorias de objetos como entidades semânticas que regem a descrição semântica global de objetos. Especificamente, pretende-se seguir uma abordagem que emula o comportamento humano nas tarefas de descrição global de objetos.

O ser humano, geralmente, usa os processos cognitivos de *generalização* e *discriminação* para construir descrições globais de objetos. Com essa estratégia, a abordagem de descrição global humana descreve o objeto destacando as características que o tornam único dentro de uma categoria específica. Um exemplo de descrição humana: *o dálmata é um cão* (habilidade de generalização para reconhecer nas características do objeto o significado semântico central da categoria cão) *branco com exclusivas manchas de cor preta ou fígado* (habilidade de discriminação para detectar as características distintivas do objeto dentro da categoria cão). Observa-se que com essa abordagem o ser humano usa as mesmas características extraídas do objeto para, primeiramente, categorizá-lo e, seguidamente, descrevê-lo. Por conseguinte, essa abordagem humana de descrição semântica global de objetos constitui a motivação principal desta pesquisa.

Mas, como modelar uma descrição global da imagem do objeto com um comportamento semelhante à abordagem humana de descrição semântica de objetos? Como descrever globalmente objetos com base nas mesmas características aprendidas para categorizá-los? Uma vez reconhecida a categoria à qual o objeto pertence, quais características o distinguem semanticamente dos outros membros dessa categoria?

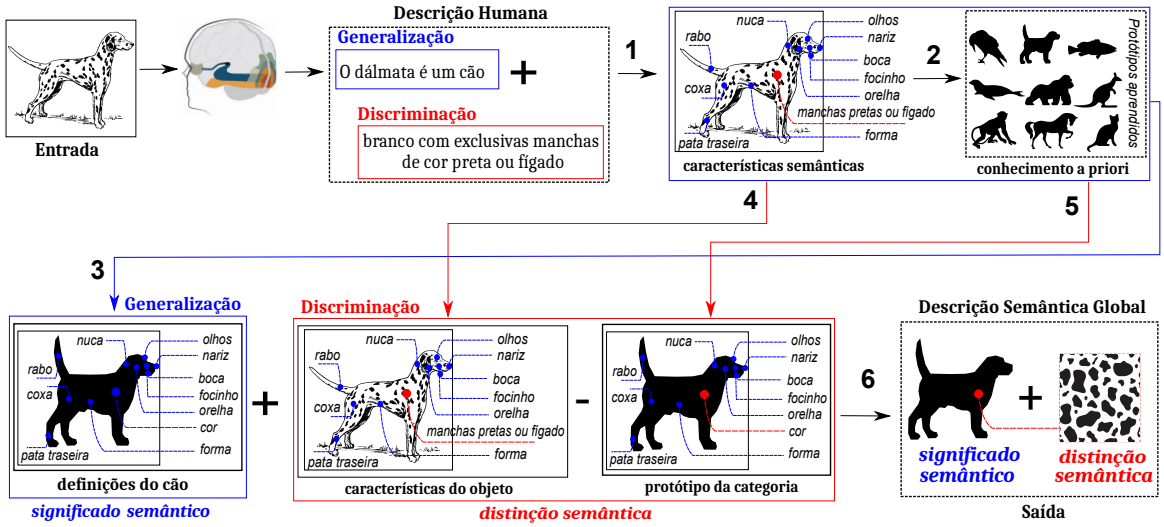


Figura 1.2: *Arcabouço da hipótese de descrição semântica de objetos baseada em protótipos proposta.* O sistema visual humano observa um objeto e pode construir uma descrição semântica que destaca as características mais distintivas do objeto dentro da categoria. A Figura apresenta um modelo de descrição baseado em protótipos que visa simular esse comportamento humano através do fluxo de processos 1-6: *1)* extração de características; *2)* reconhecimento das características do objeto; *3)* categorização; *4)* características do objeto; *5)* significado semântico central da categoria (protótipo da categoria); *6)* Descrição Semântica Global do objeto. Fonte: Elaborado pelo autor.

A Figura 1.2 apresenta o fluxo de processos da hipótese proposta para a descrição semântica de imagens de objetos usando protótipos, a qual simula a abordagem humana de descrição global de objetos. A Descrição Semântica Global do objeto (resultado do processo 6) é uma representação que encapsula o *significado semântico* do objeto e destaca as características que o distinguem dentro da categoria (*distinção semântica* em relação ao protótipo da categoria). Observe-se que: i) a hipótese apresentada assume que é possível representar o protótipo semântico da categoria; ii) a descrição semântica do objeto precisa do reconhecimento do objeto antes de descrevê-lo.

Este trabalho inspira-se na habilidade que possui o sistema de percepção visual humano para usar as mesmas características extraídas nas tarefas de reconhecimento e descrição de objetos. Não obstante, a motivação principal que torna inovadora esta pesquisa, fundamenta-se na utilização dos conceitos da semântica cognitiva e da psicolinguística para a simulação do comportamento da *memória semântica* nas tarefas de reconhecimento e descrição semântica de objetos. Especificamente, propõe-se obter um modelo de representação semântica das categorias de objetos e de descrição semântica de objetos, baseado nos fundamentos teóricos relacionados com a *representação do significado semântico* e com a *aprendizagem de conceitos visuais* da Teoria

dos Protótipos.

1.1 Definição do Problema

Como descrever semanticamente os objetos? Como modelar um sistema computacional que possua a habilidade de descrever semanticamente os objetos usando as mesmas características aprendidas para classificá-los?

1.2 Objetivo Geral

Modelar um descritor semântico global fundamentado na Teoria dos Protótipos que inclua o *protótipo da categoria* na descrição semântica do objeto.

1.2.1 Objetivos Específicos

1. Modelar matematicamente a representação semântica do protótipo das categorias de objetos, com o propósito de *codificar* e *armazenar* o *significado semântico central* (protótipo) das categorias de objetos;

2. *Reconhecer* o significado semântico central da categoria (protótipo da categoria) a partir das características extraídas de um objeto;

3. Construir um modelo de descrição semântico global baseado em protótipos que encapsule o significado semântico central da categoria na representação semântica global do objeto;

4. Avaliar o desempenho da representação do Descritor Semântico Global proposto em tarefas de Visão Computacional como agrupamento, classificação de imagens, etc;

5. Desenvolver uma camada CNN para introduzir o *conceito de categorização baseado em protótipos* nos modelos CNNs de classificação.

1.3 Contribuições

1. Uma nova abordagem fundamentada na Teoria dos Protótipos para a descrição semântica global de imagens de objetos. A abordagem de descrição semântica proposta descreve o objeto mediante representações semânticas interpretáveis

construídas seguindo os processos de *codificação*, *armazenamento* e *recuperação* do *significado semântico central* (protótipo) da categoria.

2. Um modelo matemático para representar e construir o protótipo semântico das categorias de objetos usando qualquer modelo de classificação. O modelo proposto transfere o conhecimento aprendido pelos modelos de classificação pré-treinados para uma estrutura semântica (protótipo) que tenta representar o *significado semântico central* das categorias apreendidas;

3. Uma métrica de distância semântica no domínio das características de objetos, a qual pode ser entendida como uma estimativa da tipicidade do objeto dentro da categoria. A métrica de distância proposta permite a organização prototípica dos elementos dentro da categoria;

4. Um método simples que reduz a dimensionalidade da representação semântica global do objeto, e constrói uma assinatura de pequena dimensionalidade que encapsula a informação semântica do objeto.

5. Um novo Descritor Semântico Global de objetos baseado nos fundamentos teóricos da Teoria dos Protótipos. O modelo de descrição semântica global proposto constitui um ponto de partida para introduzir os fundamentos teóricos relacionados à *representação do significado semântico* e à *aprendizagem de conceitos visuais* da Teoria dos Protótipos na família de Descritores-CNN;

6. Um método de classificação que introduz o conceito de categorização da Teoria dos Protótipos nos modelos CNNs de classificação. Foi proposta a Camada de Similaridade Prototípica (PS-Layer) que mantém os protótipos calculados como conhecimento prévio em seus neurônios e produz como saída do neurônio a distância semântica proposta. A saída da camada PS-Layer pode ser interpretada como uma distribuição de probabilidades da similaridade entre o objeto e os protótipos aprendidos.

1.4 Publicações

Os resultados deste trabalho e as contribuições foram submetidas para publicação:

- Pino, O.; Nascimento, E.; Campos, M. (2019). Prototypicality effects in global semantic description of objects. Em *Proceedings of the IEEE Winter Conference on the Applications of Computer Vision (WACV)*, pp. 1233-1242. ISSN 1550-5790.

- Pino, O.; Nascimento, E.; Campos, M. (2019). Global Semantic Description of Objects based on Prototype Theory. *IEEE Transactions on Image Processing (TIP)*. (Submetido)

1.5 Organização

O presente documento está organizado da forma que segue:

- **Capítulo 1: Introdução**. Apresentação do problema, motivação da pesquisa, e os objetivos estabelecidos.
- **Capítulo 2: Trabalhos Relacionados**. Discussão das principais pesquisas da literatura relacionadas com a descrição de características da imagem.
- **Capítulo 3: Teoria dos Protótipos**. Discussão dos principais fundamentos da Teoria dos Protótipos na literatura que sustentam a metodologia proposta.
- **Capítulo 4: Metodologia Geral**. Apresentação da metodologia geral da pesquisa. Apresenta-se o fluxo de processos que devem ser desenvolvidos na construção do modelo de descrição semântico global de objetos baseado em protótipos.
- **Capítulo 5: Modelo Computacional do Protótipo**. Apresentação das definições e do modelo matemático proposto para encapsular o protótipo semântico da categoria. Apresentação da métrica de distância semântica proposta para quantificar a similaridade do objeto com o protótipo da categoria. Apresentação dos resultados experimentais obtidos na avaliação do Modelo Computacional do Protótipo proposto.
- **Capítulo 6: Descritor Semântico Global**. Apresentação da metodologia do modelo de descrição semântica global de imagens de objetos baseado no Modelo Computacional do Protótipo proposto. Descrição do método proposto para a redução da dimensionalidade da representação global semântica do objeto. Apresentação dos resultados experimentais obtidos nessa parte da metodologia.
- **Capítulo 7: Classificação semântica baseada em Protótipos**. Apresenta-se a abordagem proposta que torna a tarefa de reconhecimento do protótipo uma tarefa de classificação semântica de objetos. Apresentam-se os resultados experimentais obtidos nessa parte da metodologia.
- **Capítulo 8: Conclusões**.

Capítulo 2

Trabalhos Relacionados

A detecção e a descrição de características constituem tarefas relevantes e tem sido objeto de pesquisa em Visão Computacional. Os avanços tecnológicos e as técnicas computacionais atuais posicionam a descrição de características na intersecção de várias áreas como a Visão Computacional. O advento das Redes Neuronais Convolucionais (CNNs) permitiu, pela primeira vez, construir modelos de reconhecimento, de detecção e de memória com desempenho próximo ao ser humano. Os modelos de reconhecimento visual e extração de conhecimento baseados em CNN ofereceram soluções mais robustas em comparação com outras técnicas anteriores, uma vez que as tarefas de extração e descrição de características foram altamente influenciadas pelo uso dessas técnicas de aprendizagem profunda. Neste capítulo são apresentados os principais trabalhos relacionados com a detecção e a descrição de características da imagem, com ênfase nas abordagens de descrição semântica das características da imagem.

2.1 Detecção e Descrição de Características

As caraterísticas da imagem constituem pequenas regiões (*patches*)[1] ou propriedades úteis para estabelecerem similitudes entre imagens. Indistintamente são usados vários conceitos para definir essas caraterísticas, mas é válido esclarecer as diferenças e as particularidades conceituais neste contexto.

Sob a abordagem da aprendizagem de máquina e reconhecimento de padrões, uma caraterística (*feature*) é definida como uma propriedade individual mensurável ou uma caraterística de um fenômeno observado (Bishop, 2006). As métricas para escolherem as características da imagem evoluíram, mas sempre com a premissa de que fossem

[1]Doravante será usado o termo em inglês *patch* ou *patches* para fazer referência a pequenas regiões de uma imagem.

Tabela 2.1: Propriedades de uma boa característica. Fonte: Adaptado de Szeliski, 2010, p. 207.

Propriedade	*Descrição*
Repetitividade/Precisão (*Repeatability/Precision*)	A mesma caraterística pode ser encontrada em várias imagens apesar das transformações geométricas e fotométricas.
Saliência/Correspondência (*Saliency/Matchability*)	Cada característica possui uma descrição distinta ou típica.
Compacidade e Eficiência (*Compactness and efficiency*)	Representar muito menos caraterísticas (*features*) que píxeis na imagem.
Localidade (*Locality*)	A caraterística ocupa uma área relativamente pequena da imagem; esse atributo permite ser robusto ante a desordem (*clutter*) e a oclusão (*occlusion*).

informativas, discriminativas e independentes. Esses últimos atributos são essenciais para eficiência de outros algoritmos como os algoritmos de classificação, de regressão e de reconhecimento de padrões.

Nas áreas de Visão Computacional e do Processamento de Imagem, uma caraterística constitui uma informação relevante na imagem para resolver determinada tarefa computacional. Uma característica é definida como uma região ou propriedade na imagem com certa importância dado determinado contexto ou métrica (Forsyth & Ponce, 2011). Segundo Szeliski (2010) uma boa caraterística deve possuir quatro propriedades como se apresenta na Tabela 2.1.

Uma característica da imagem é composta, geralmente, por um par: um ponto-chave (*keypoint*) e um descritor da característica que a descreve (Lowe, 2004). O ponto-chave normalmente contém a posição do patch 2D entre outras informações como a escala e a orientação da caraterística correspondente. Aliás, o descritor encapsula a descrição visual da caraterística ou *patch* e a utiliza na comparação de similitudes entre as características da imagem (correspondência).

Um ponto-chave é uma característica local que se refere a recursos como quinas, arestas ou padrões que podem ser encontrados repetidamente com alta probabilidade e que desempenham um papel fundamental em muitas aplicações de Visão Computacional. Em dependência da taxonomia da informação representada nos descritores de pontos-chaves, foram desenvolvidas ao longo dos anos numerosas pesquisas e linhas de "famílias" separadas de descritores de características. Em muitos casos, as comunidades de pesquisa para cada família trabalham em problemas diferentes, e existe pouco interesse mútuo (Krig, 2014). A Figura 2.1 mostra a taxonomia das famílias de des-

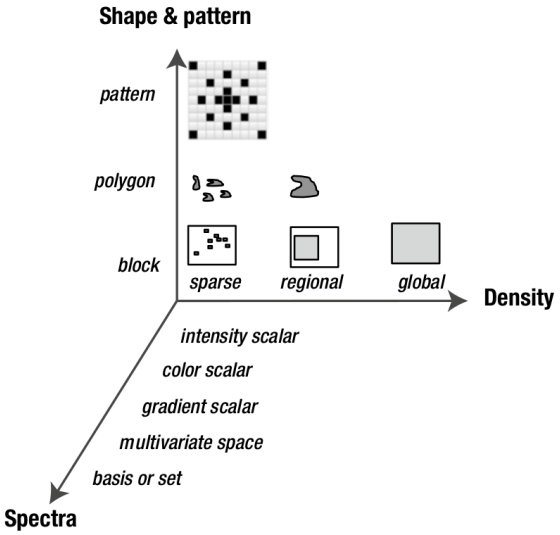


Figura 2.1: *Taxonomia das famílias de descritores de características.* Os eixos representam: *i)* a densidade da característica (*Density*): global, regional e esparso local; *ii)* forma e padrão (*Shape and pattern*) dos píxeis usados na construção do descritor: retângulos, polígonos e padrões de amostragem esparsos; *iii*) espectros (*Spectra*) das informações contidas na representação da característica. Fonte: Krig, 2014, p. 192.

critores de características (Krig, 2014). As dimensões com três eixos (forma e padrão, espectros, e densidade) simplificam a análise dessas famílias.

Todos os descritores de características são projetados para fornecerem representações discriminativas das características salientes na imagem, visando ser robustos a transformações como ponto de vista, escala, mudanças de iluminação, etc. De maneira geral, um descritor de caraterísticas deve cumprir com duas propriedades principais: invariante e distintivo (Forsyth & Ponce, 2011; Szeliski, 2010). A propriedade de invariância refere-se à habilidade de descrever (quase) da mesma maneira uma determinada caraterística em imagens diferentes. Ser distintivo alude à propriedade, da métrica do descritor, de ser caraterístico e particular na representação daquela caraterística. Embora sejam esses os atributos principais, um bom descritor de caraterísticas e pontos-chave (em dependência da tarefa) deve possuir outros atributos (Ver Anexo A).

A propriedade de invariância às transformações geométricas na imagem (Ver exemplos na Figura 2.2), de iluminação, de escala, etc. são desejadas pelas infinitas aplicações práticas. Alguns algoritmos na extração e descrição de características garantem essas invariâncias separadamente. Por exemplo, Lowe (2004) desenvolveu com SIFT (*Scale Invariant Feature Transform*) um detector-descritor de características: o detector é invariante às transformações de escala e de rotação, enquanto o descritor adiciona invariância às mudanças de iluminação e invariância parcial às distorções afins

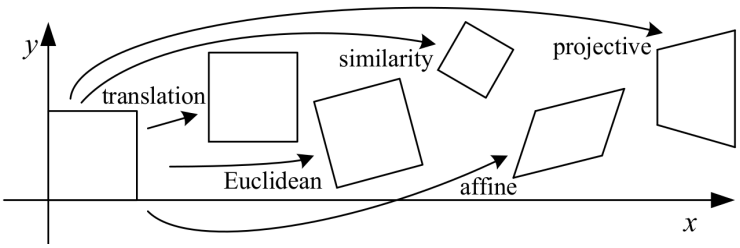


Figura 2.2: Conjunto básico de transformações planares no espaço 2D. Fonte: Szeliski, 2010, p. 36.

(*affine*). Outros métodos combinaram detectores diferentes com o descritor SIFT; tornando o descritor a parte mais usada do algoritmo de Lowe (2004).

A descrição de características atingiu uma alta maturidade com a introdução da proposta do SIFT (Lowe, 2004). O cálculo desse descritor a partir de histogramas locais de orientações do gradiente (Ver Figura 2.3) criou as bases para que outras tentativas como SURF (Bay et al., 2008) usaram as representações da imagem integral para acelerar o processamento. Uma abordagem semelhante foi proposta por Tola et al. (2008) com DAISY, que se baseou em mapas convolvidos de gradientes orientados para aproximarem os histogramas; essa abordagem melhorou o desempenho computacional na extração de descritores densos.

As famílias de descritores podem ser classificadas segundo a taxonomia dos descritores apresentada na Figura 2.1 e também, segundo a evolução do processo de extração e descrição de características como segue abaixo:

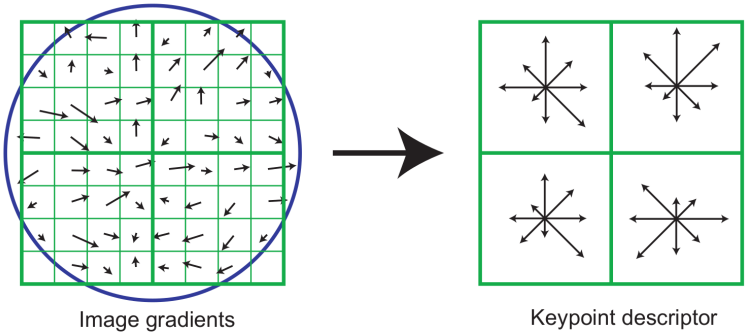


Figura 2.3: Visão geral da construção do descritor de características SIFT. O descritor é construído calculando, primeiramente, a magnitude e a orientação do gradiente da imagem na região em torno à localização do ponto-chave (esquerda). Os gradientes são ponderados por uma janela gaussiana, indicada pelo círculo superposto. As amostras são acumuladas em histogramas de orientação que resumem os conteúdos em subregiões 4x4. O comprimento de cada seta corresponde à soma das magnitudes do gradiente, perto dessa direção dentro da região. Fonte: Lowe, 2004, p. 15.

- Descritores Locais Binários (*Local Binary Descriptors*): Exemplos dessa família são LBP (Ojala et al., 1996), FREAK (Alahi et al., 2012), ORB (Rublee et al., 2011), BRISK (Leutenegger et al., 2011), Census (Zabih & Woodfill, 1994).

- Descritores de Espectros (*Spectra Descriptors*): SIFT e as suas variantes, SURF, e CenSurE (Agrawal et al., 2008) constituem os exemplos mais representativos dessa família.

- Descritores Espaciais Básicos (*Basis Space Descriptors*): Descritores baseados na Transformada de Fourier constituem os máximos exemplos dessa família.

- Descritores de Forma Poligonal (*Polygon Shape Descriptors*): Esses usam a forma de objetos medidos por métricas estatísticas, tais como área, perímetro e centroide. O MSER (Donoser & Bischof, 2006) pode ser considerado como o fundamento dos descritores de forma.

- Descritores Globais (*Global Descriptors*): Os descritores globais descrevem a imagem como um todo para generalizar todo o objeto. Exemplos de representações globais da imagem são: GIST (Oliva & Torralba, 2001), LBP (Ojala et al., 2002), HOG (Dalal & Triggs, 2005a), Color64 (Li, 2007), Color_Hist (Song et al., 2004), *Image Moments* (Hu, 1962), Haralick (Haralick et al., 1973), etc.

Apesar de que essas abordagens terem sido extremamente bem-sucedidas, desenvolvimentos recentes na construção de descritores locais na imagem estão sendo desviados desses enfoques cuidadosamente projetados (Lowe, 2004; Bay et al., 2008; Tola et al., 2008; Rublee et al., 2011; Mainali et al., 2014); e estão aproximando-se à aprendizagem de características em grandes volumes de dados. Essa linha de trabalho inclui técnicas não supervisionadas baseadas em *Hashing* (Ambai & Yoshida, 2011), bem como abordagens supervisionadas usando Análise Discriminativo Linear (Brown et al., 2011; Gong et al., 2013; Strecha et al., 2012), *Boosting* (Trzcinski et al., 2013), Algoritmo Genético (Perez & Olague, 2013) e Otimização Convexa (Simonyan et al., 2014). Essas abordagens demonstraram que aqueles descritores artesanais agora podem ser superados em desempenho por novos descritores aprendidos.

Historicamente a escolha das métricas de características foi limitada àquelas que podiam ser calculadas nesse momento devido às limitações em poder computacional, memória e tecnologia daqueles sensores. Com o desenvolvimento tecnológico atual, as métricas tornaram-se mais complexas para seu cálculo, requerendo maior poder de processamento e memória. As imagens evoluíram em multimodais, combinando: intensidade, cor, espetros múltiplos, informação do sensor de profundidade, maior velocidade

de fotogramas e maior precisão e exatidão nas coordenadas X, Y, Z. O aumento na largura de banda da memória e no desempenho do cálculo, originaram novas formas de descreverem características e novas métricas para analisar o desempenho.

Nesse sentido, uma tendência recente consiste na extração de características locais (extraídas diretamente dos *patches* brutos da imagem) e características globais (extraídas de regiões ou de toda a imagem) mediante o uso das CNNs treinadas em grandes volumes de dados. Essa nova abordagem define o advento de uma nova família dentro da taxonomia apresentada por Krig (2014) denominada descritores baseados/aprendidos com CNNs (*CNN - based descriptors*).

2.2 Descritores Aprendidos usando CNNs

O advento das CNNs possibilitou a obtenção de modelos de reconhecimento da informação visual contida na imagem que demostraram excelente desempenho em tarefas de classificação (Simonyan & Zisserman, 2014; Chollet, 2016; He et al., 2016; Szegedy et al., 2016; Howard et al., 2017; Szegedy et al., 2017). O sucesso das CNNs desencadeou a tendência do processamento da imagem usando essas técnicas de aprendizagem profunda em diversas tarefas de Visão Computacional, tais como descrição de imagens em linguagem natural (Karpathy & Fei-Fei, 2015), recuperação da imagem baseada em conteúdo (Zhu et al., 2017), classificação da cena (Nogueira et al., 2017), etc. (Vide o Apêndice A para um breve resumo sobre a evolução, estrutura, conceitos principais e modelos relevantes das CNNs).

Em consequência, mesmo quando algoritmos como SIFT (Lowe, 2004), SURF (Bay et al., 2008) e DAISY (Tola et al., 2008) corroboraram a maturidade que tinha alcançado a descrição de características da imagem, várias abordagens CNNs foram desenvolvidas nos últimos anos para a aprendizagem de características discriminatórias em grandes conjuntos de dados, conseguindo ultrapassar em desempenho essas abordagens bem estruturadas.

Alguns enfoques como os propostos por Donahue et al. (2014); Fischer et al. (2014); Gong et al. (2014); Long et al. (2014) extraíram ativações imediatas como descritor, mostrando serem eficazes para a correspondência no nível de *patch*. Outros métodos aprenderam diretamente uma medida de similaridade para comparar *patches* usando uma rede de similaridade convolucional (Han et al., 2015; Simo-Serra et al., 2015; Zagoruyko & Komodakis, 2015; Yi et al., 2016).

Os estudos de Jahrer et al. (2008) e Osendorfer et al. (2013) constituem os primeiros antecedentes focados na descrição de caraterísticas mediante a aprendizagem

profunda. Os resultados experimentais obtidos pelos autores não foram muito concludentes, mas deixaram questões abertas como: quais arquiteturas de rede seriam as mais apropriadas e, dependendo da aplicação, quais esquemas de treinamento usar.

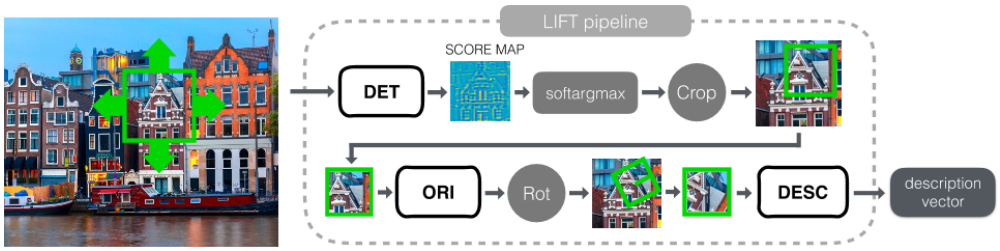
Baseado no sucesso de AlexNet, Fischer et al. (2014) analisaram o desempenho dos descritores convolucionais, construídos usando a rede AlexNet, no banco de dados de Mikolajczyk & Schmid (2005). Os resultados conseguidos pelos autores mostraram que esses descritores convolucionais aprendidos superavam o algoritmo SIFT em desempenho. Outra contribuição interessante de Fischer et al. (2014) residiu na abordagem de treinamento não supervisada usada para aprenderem os seus descritores.

Aliás, os trabalhos de Han et al. (2015); Zagoruyko & Komodakis (2015); Zbontar & LeCun (2015) usaram as redes siamesas para aprendizagem de descritores e para a aprendizagem da métrica de similitude. A rede Siamesa constitui um tipo de rede cujo funcionamento é semelhante à ideia de calcular um descritor (Bromley et al., 1994; Chopra et al., 2005). Existem dois ramos da mesma rede que compartilham a mesma arquitetura e os mesmos pesos. No funcionamento, cada ramo utiliza como entrada um *patch* e processa, nas camadas interiores, o descritor correspondente, usando uma rede superior como função de similitude.

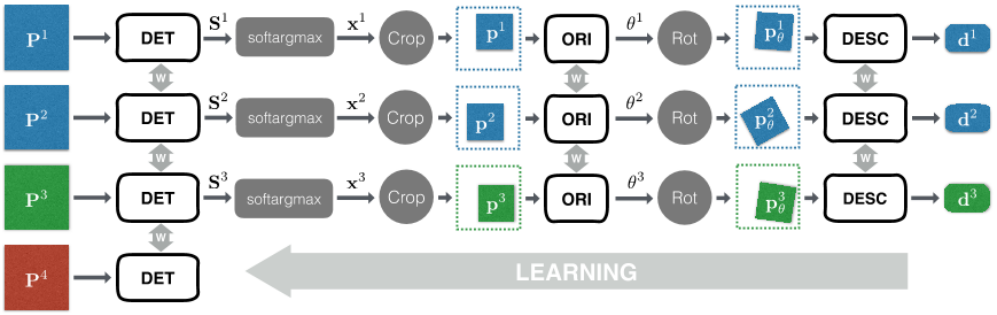
Especificamente, Han et al. (2015) propuseram a rede convolucional profunda MatchNet que possui, além de uma arquitetura siamesa, uma rede totalmente conectada para aprender a função de comparação dos descritores. Por outra parte, DeepCompare (Zagoruyko & Komodakis, 2015) possui uma arquitetura semelhante, mas os autores adicionaram uma rede que se concentrava apenas no centro da imagem.

Zbontar & LeCun (2015) propuseram uma abordagem de comparação de *patches* para calcular o custo da correspondência no problema clássico de correspondência estéreo. Os autores obtiveram resultados de última geração e mostraram o melhor desempenho no conjunto de dados KITTI (Geiger et al., 2013). As abordagens anteriores (Han et al., 2015; Zagoruyko & Komodakis, 2015; Zbontar & LeCun, 2015) dependem de redes maiores e não necessitam de representações compactas e discriminativas.

Nesse sentido, Simo-Serra et al. (2015) propuseram uma abordagem baseada em representações compactas e discriminativas. Os autores apoiaram-se na mineração negativa (*hard negative mining*) para aprender descritores compactos de dimensionalidade 128-D. Os descritores estavam constituídos por ativações imediatas e eram comparados usando a distância euclidiana como função de similitude. Os autores demonstraram um aumento significativo no desempenho com relação aos estudos anteriores. Os descritores aprendidos por Simo-Serra et al. (2015) possuem um bom funcionamento na presença de mudanças de escala e rotação, de transformação de perspectiva, de deformação não rígida e em mudanças de iluminação.



(a) Visão Geral da arquitetura de LIFT.



(b) Detalhe da arquitetura siamesa usada para o treinamento.

Figura 2.4: Arquitetura geral do algoritmo LIFT. A arquitetura de rede profunda aprende cada uma das tarefas envolvidas no gerenciamento de características: detecção, estimativa de orientação e descrição de características. Fonte: Yi et al., 2016.

A aprendizagem de descritores que usam como função de similitude a distância Euclidiana possui uma ampla gama de aplicação com relação àqueles descritores que exigem uma métrica aprendida. Essa característica mostrou um excelente desempenho na abordagem de Simo-Serra et al. (2015) e constituiu o fundamento para os estudos de Yi et al. (2016).

A arquitetura de LIFT (*Scale Invariant Feature Transform*) proposta por Yi et al. (2016) pode ser considerada como a versão de aprendizagem profunda de SIFT. Os autores apresentaram uma arquitetura de rede profunda que aprende cada uma das tarefas envolvidas no gerenciamento de características: detecção, estimativa de orientação e descrição de características (Figura 2.4a). Nesse trabalho foram apresentadas essas três tarefas de forma unificada, preservando a diferenciação da rede de ponta a ponta, uma abordagem sem precedentes. Os resultados alcançados superaram os métodos anteriores em vários bancos de dados (Strecha et al., 2008; Aanæs et al., 2012; Verdie et al., 2015) sem necessidade de um novo treinamento.

Na estrutura detalhada de LIFT (Figura 2.4) cada componente individual possui objetivos diferentes que impedem o treinamento desde zero da arquitetura completa. Yi et al. (2016) treinaram o descritor, e com esse resultado treinaram o Estimador de Orientação. Com esses subcomponentes aprendidos, o Descritor e o Estimador de Ori-

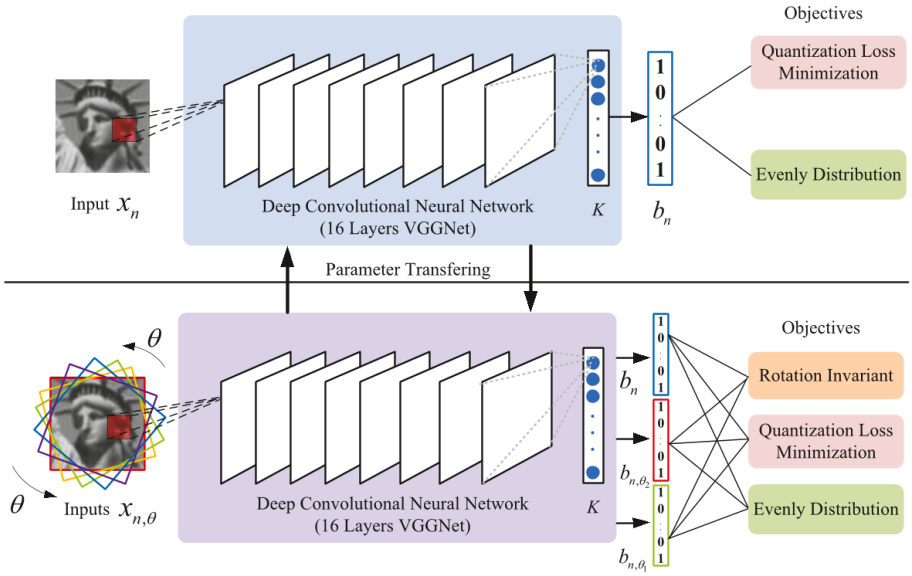


Figura 2.5: Visão geral da arquitetura de DeepBit. Os códigos binários do descritor são construídos usando as características extraídas com o modelo da rede VGG. Fonte: Lin et al., 2016b.

entação treinaram o Detector, logrando uma diferenciação por toda a rede. Na etapa de teste os autores desacoplaram o Detector para usá-lo na imagem, deixando apenas o Estimador de Orientação e o Descritor para o processamento dos pontos-chave extraídos. Os resultados experimentais de LIFT demonstraram que a abordagem integrada usada possui um excelente desempenho, alcançando resultados que constituem o estado da arte atual.

Outro resultado relevante, que constitui o estado da arte atual na representação binaria de características, foi obtido por Lin et al. (2016b). A proposta de DeepBit (Lin et al., 2016b) usa uma abordagem não supervisionada na aprendizagem profunda de um descritor binário compacto para correspondência eficiente de objetos visuais. Os autores, na camada superior da rede apresentada, usaram três critérios para projetar os códigos binários do descritor: quantificação de perda mínima, códigos uniformemente distribuídos e bits não correlacionados.

A Figura 2.5 apresenta a visão geral da arquitetura de DeepBit. Os autores projetaram seus critérios —na construção dos códigos binários do descritor— usando as características extraídas (neurônios na camada superior) com o modelo da rede VGG (Simonyan & Zisserman, 2014). O processo de treinamento foi realizado em duas etapas. Na primeira etapa (linha superior da Figura 2.5) foram aprendidos os parâmetros da rede tentando conseguir uma quantificação de perda mínima enquanto geravam códigos binários de maneira uniforme. No segundo estágio do treinamento (linha inferior da Figura 2.5) os dados são aumentados com rotações diferentes visando

obter invariância às transformações de rotação. Os parâmetros compartilhados da rede são atualizados minimizando a distância entre os descritores binários obtidos da imagem de referência e da imagem rotacionada. No algoritmo de Lin et al. (2016b) essas etapas alternativas são repetidas até que seja satisfeito certo critério de parada estabelecido pelos autores.

Os autores demonstraram que os descritores binários extraídos com DeepBit possuem as características de ser compactos e discriminativos. Os resultados experimentais alcançados com o algoritmo em três tarefas diferentes de análise visual — correspondência de imagens, recuperação de imagens e reconhecimento de objetos— demonstram claramente a eficácia dessa abordagem. Como característica interessante, a abordagem de Lin et al. (2016b) —em comparação com descritores binários supervisionados— não requer de dados rotulados durante a aprendizagem, uma habilidade prática para aplicações no mundo real. Os resultados em três bancos de dados (Nilsback & Zisserman, 2006; Krizhevsky & Hinton, 2009; Brown et al., 2011) demonstraram que o método proposto pelos autores atinge o melhor desempenho em comparação com outros descritores binários de características.

2.3 Descritores Semânticos

As abordagens clássicas de correspondência (Okutomi & Kanade, 1993; Szeliski, 2006; Scharstein & Szeliski, 2002), problema central em muitas tarefas de Visão Computacional, foram desenhadas para encontrar correspondência entre imagens ou cenas que contêm os mesmos objetos, mas com mudanças moderadas com relação aos pontos de vistas e às transformações geométricas. Essas técnicas assumem que as imagens que serão registradas contêm os mesmos valores de pixel depois de aplicar as transformações geométricas. O processo de correspondência tornou-se mais complicado em cenas dinâmicas como as sequências de *frames* em vídeos. O Fluxo Óptico (Horn & Schunck, 1981; Bruhn et al., 2005) predominou como paradigma na resolução desse tipo de problema.

A abordagem de encontrar correspondências entre cenas diferentes, mas que compartilham características similares ou semanticamente relacionadas, aumentou a complexidade da tarefa de correspondência clássica. Esse tipo de correspondência "semântica" constitui um desafio pela presença de variações importantes na imagem: oclusão, desordem de fundo, mudanças de pose e iluminação na captura de imagens e vídeos, presença de objetos diferentes mas semanticamente iguais, etc.

Inspirado pelos métodos de Fluxo Óptico, Liu et al. (2011) (SIFT Flow) introdu-

ziram uma abordagem de correspondência a nível semântico para resolver o problema de alinhamento da cena (*scene aligment*). Essa abordagem criou uma família de métodos de fluxo semântico (*semantic flow*) (Liu et al., 2011; Kim et al., 2013; Trulls et al., 2013; Yang et al., 2014; Qiu et al., 2014; Bristow et al., 2015) como uma solução ante o alto grau de variação que inclui o desafio da correspondência semântica.

Todas essas abordagens de correspondência semântica usaram descritores de características e/ou esquemas de melhoras de desempenho (*optimization schemes*) que, mesmo com um correto funcionamento, utilizavam um regularizador espacial para garantir a fluidez do fluxo baseado em características artesanais ou aprendidas como SIFT (Lowe, 2004), HOG (Dalal & Triggs, 2005b) e DAISY (Tola et al., 2008). Fato que constituiu uma limitação devido ao surgimento de métodos aprendidos que superavam em desempenho esses descritores artesanais de características.

A propriedade de invariância nas características (*features*) é essencial para a efetividade da tarefa de correspondência. As CNNs garantem –por construção, estrutura e natureza de as suas operações– algumas invariâncias que motivaram seu uso em tarefas de correspondência. As primeiras abordagens foram usadas na resolução do problema clássico de correspondência, como fluxo óptico e correspondência estérea, para aprender descritores de características (Zagoruyko & Komodakis, 2015; Zbontar & LeCun, 2015, 2016) ou funções de similaridade (Han et al., 2015; Zagoruyko & Komodakis, 2015; Zbontar & LeCun, 2015).

Dosovitskiy et al. (2015) propuseram o FlowNet usando um esquema de ponta a ponta para aprender o fluxo óptico com um conjunto de dados sintéticos. A rede aprendida apresentada pelos autores conseguiu funcionar corretamente em vários bancos de dados existentes, mas os resultados alcançados não superaram os métodos tradicionais de fluxo óptico como EpicFlow (Revaud et al., 2015).

Nesse sentido, outras abordagens de aprendizagem supervisionada apareceram. Algumas propostas recentes como Han et al. (2015); Zagoruyko & Komodakis (2015); Zbontar & LeCun (2015) e Zbontar & LeCun (2016) usam supervisão de cenas 3D reconstruídas e pares estéreos (*stereo pairs*). De especial importância são os trabalhos de Zbontar & LeCun (2015, 2016) na proposta de MC-CNN (Zbontar & LeCun, 2015) e a sua extensão (Zbontar & LeCun, 2016). Os autores, primeiramente, treinaram os modelos CNN propostos para prever a exatidão com que dois pontos da imagem correspondem; e depois usaram essas informações para calcular o custo da correspondência estéreo.

As características aprendidas com CNNs tornaram-se cada vez mais usadas para a estimativa da correspondência. As abordagens de Agrawal et al. (2015); Han et al. (2015); Zagoruyko & Komodakis (2015); Simo-Serra et al. (2015); Yi et al. (2016) e

Lin et al. (2016b) aprenderam uma função de similaridade para *patches*, diretamente a partir de imagens de uma cena 3D. Esses trabalhos mostraram uma alta precisão de localização dos pontos correspondentes e de invariância em pequenas transformações geométricas e fotométricas.

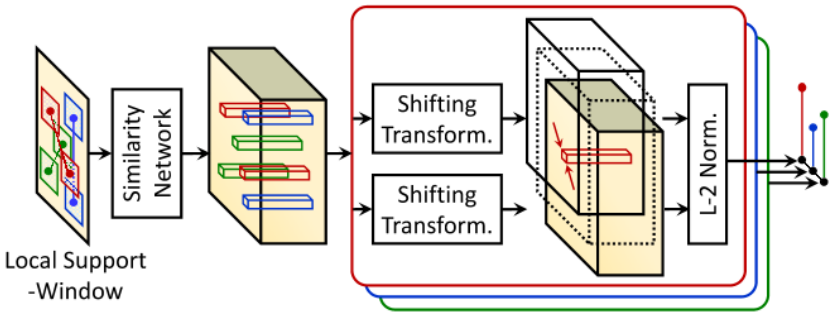
Os enfoques anteriores estão inerentemente limitados às imagens coincidentes do mesmo objeto físico/cena, pelo que não possuem a habilidade de lidar com grandes variações ou mudanças nas imagens. Ante esse tipo de problema foram desenvolvidas novas abordagens. A correspondência semântica usando CNN surgiu para conseguir uma maior invariância e para lidar com as diferenças de aparência mais substanciais.

A resolução da tarefa de correspondência semântica sob o enfoque da aprendizagem profunda possui o problema da inexistência de conjuntos de dados rotulados disponíveis para essa tarefa. Nesse sentido, Long et al. (2014) usaram caraterísticas aprendidas para tarefas de classificação (ImageNet) alcançando um desempenho comparável a SIFT Flow. Zhou et al. (2016) aproveitaram os modelos 3D e utilizaram a consistência do fluxo entre modelos 3D e imagens 2D como sinal de supervisão para treinar os modelos CNN. Outra abordagem para a geração de dados (*ground truth*) consistiu em aumentar diretamente os dados criando anotações densas mediante deformação (Ham et al., 2016; Kanazawa et al., 2016) de anotações esparsas de pontos-chave.

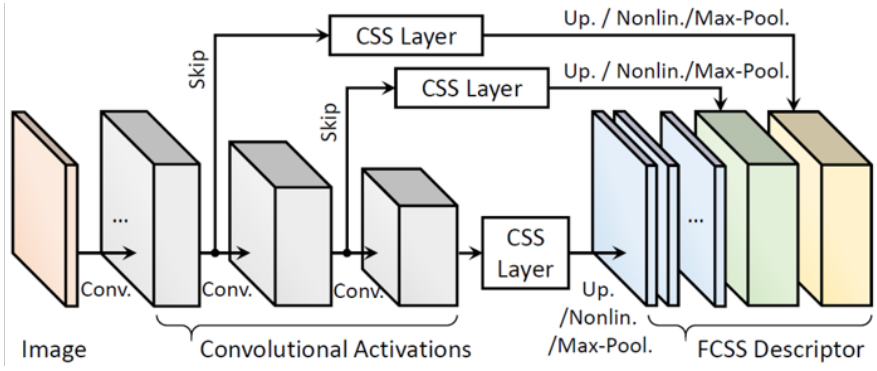
A rede de correspondência universal (*Universal Correspondence Network (UCN)*) proposta por Choy et al. (2016) aprende a correspondência semântica usando uma arquitetura semelhante à abordagem de Zbontar & LeCun (2016), mas adiciona redes de transformações espaciais convolucionais (Jaderberg et al., 2015) para maior robustez nas mudanças de rotação e de escala. Como esses métodos formam as estruturas das CNNs combinando redes convolucionais existentes, os autores propõem uma abordagem custo-benefício entre a invariância de aparência e entre a precisão de localização; essa abordagem possui limitações inerentes à correspondência semântica.

Kim et al. (2017) apresentou um descritor baseado em CNN que possui a habilidade de ser intrinsecamente insensível às variações de aparência dentro da classe, mantendo a capacidade de localização precisa. O descritor apresentado pelos autores (Ver Figura 2.6), denominado Auto-similitude Completamente Convolucional (*Fully Convolutional Self-Similarity (FCSS)*), combina de forma robusta os pontos entre diferentes instâncias dentro da mesma classe do objeto. Os resultados experimentais de FCSS superaram o desempenho dos descritores artesanais convencionais e dos descritores anteriores baseados em CNN em vários bancos de dados de referência.

Os autores formularam FCSS utilizando a auto-similitude local (*Local Self-Similarity (LSS)*) dentro de uma rede completamente convolucional para a correspondência semântica densa. O objetivo principal desse trabalho, usando as propriedades



(a) Camada de auto-similitude convolucional.



(b) Detalhe da construção do descritor FCSS.

Figura 2.6: Arquitetura geral do descritor FCSS para correspondência semântica. A arquitetura de rede profunda projeta a medida de auto-similitude (LSS) para ser aprendida dentro de uma rede totalmente convolucional. O descritor FCSS é calculado usando as auto-similitudes convolucionais em escalas múltiplas. Fonte: Kim et al., 2017.

da LSS, visa obter que os *layouts* estruturais locais permaneçam aproximadamente iguais entre diferentes instâncias de objetos da mesma classe. Mesmo com cores diferentes, gradientes e pequenas diferenças nas posições das características, basicamente preserva-se a auto-similitude local (LSS) entre os pares de *patchs* comparados. Essa propriedade da LSS foi utilizada anteriormente para detecção de objetos não rígidos (Shechtman & Irani, 2007), recuperação de esboço (*sketch retrieval*)(Chatfield et al., 2009) e correspondência de modalidade cruzada (Kim et al., 2015). No entanto, como foi demonstrado por Kim et al. (2017), essas técnicas existentes baseadas em LSS podem ser definidas, principalmente, como artesanais, e precisam de mais robustez para capturar evidências de correspondência confiáveis em imagens semanticamente semelhantes.

Contrário das abordagens anteriores de Shechtman & Irani (2007); Chatfield et al. (2009) e Kim et al. (2015), o descritor proposto por Kim et al. (2017) projeta a medida de auto-similitude (LSS) para ser aprendida dentro de uma rede totalmente convolu-

cional. Os autores propuseram uma camada de auto-similitude convolucional (*convolutional self-similarity* - CSS) (Figura 2.6a) que codifica a estrutura LSS e possui diferenciação. A camada CSS projetada permite o treinamento de ponta a ponta, de modo que são aprendidos os padrões da amostragem de *patch* e a medida de auto-similitude. Na construção do descritor FCSS os autores processam as auto-similitudes convolucionais em escalas múltiplas (Figura 2.6b), usando as camadas salteadas (*skip layers*) propostas por Long et al. (2015) para encaminharem as ativações convolucionais intermediárias. Os resultados experimentais alcançados pelo descritor FCSS em vários bancos de dados de referência —Taniai et al. (2016), Proposal Flow (Ham et al., 2016), PASCAL dataset (Chen et al., 2014), e Caltech-101 (Fei-Fei et al., 2006)— superaram os descritores artesanais convencionais e os descritores baseados em CNN existentes até esse momento.

As abordagens que aprendem correspondência semântica (Choy et al., 2016; Zhou et al., 2016; Rocco et al., 2018) ou descritores semânticos (Kim et al., 2017) geralmente funcionam melhor que as técnicas tradicionais. Baseados nesses pressupostos, outro trabalho recente que aborda o problema de estabelecer correspondências semânticas entre imagens que possuem diferentes instâncias do mesmo objeto ou categoria de cena, foi apresentado por Han et al. (2017). Os autores apresentaram uma arquitetura de rede neural convolucional chamada SCNet. A proposta, diferente das propostas anteriores, usa como ideia principal a incorporação da consistência geométrica entre regiões ou partes do objeto no processo de aprendizagem. A SCNet usa propostas de objeto (*object proposals*) (Manen et al., 2013; Uijlings et al., 2013; Zitnick & Dollár, 2014) como ideia principal para a correspondência.

Em contraste com outras propostas, a rede SCNet aprende a correspondência semântica usando a aparência e as informações geométricas. Os resultados alcançados pelos autores, embora não melhoraram os resultados alcançados por Kim et al. (2017), demonstram claramente a eficácia da aprendizagem de correspondência geométrica para a correspondência semântica. A rede SCNet foi treinada com pares de imagens obtidas do conjunto de dados de pontos-chave PASCAL VOC 2007 (Chen et al., 2014). As avaliações comparativas realizadas por Han et al. (2017) em vários *benchmarks* tradicionais demonstraram que a abordagem proposta supera substancialmente arquiteturas de aprendizagem profundas recentes, assim como métodos anteriores baseados em características artesanais (*hand-craft*).

Rocco et al. (2018) apresentaram uma abordagem cujos resultados são o estado da arte atual em tarefas de correspondência semântica até o momento de desenvolvimento desta pesquisa. Os autores apresentaram um inovador trabalho que introduz três contribuições principais na tarefa de alinhamento semântico. Primeiro, apresenta-

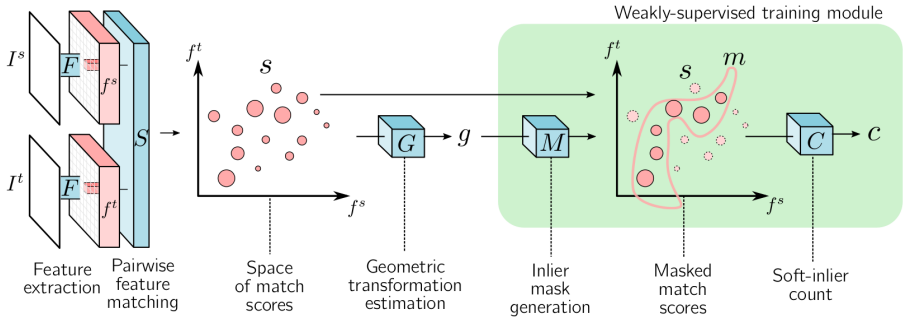


Figura 2.7: Visão Geral do modelo de alinhamento semântico de imagens inspirado pela pontuação *inlier* usada no algoritmo RANSAC. Fonte: Rocco et al., 2018.

ram uma arquitetura de rede neural convolucional para alinhamento semântico que é integralmente treinável (*end-to-end*) a partir da supervisão fraca, no nível da imagem, dos pares de imagens correspondentes. A abordagem dos autores permite que os parâmetros da rede sejam aprendidos usando a importante variação de aparência presente entre imagens diferentes (mas semanticamente relacionadas), sem a necessidade da tediosa anotação manual de correspondências no momento de treino da rede. Em segundo lugar, o principal componente da arquitetura proposta pelos autores baseia-se em um módulo diferenciável de pontuação suave de *inlier*, inspirado pelo procedimento de contagem de *inlier* do algoritmo RANSAC. A abordagem dos autores calcula a qualidade de alinhamento com base apenas em correspondências geometricamente consistentes, reduzindo o efeito de desordem de fundo.

A Figura 2.7 apresenta a abordagem proposta pelos autores. Duas imagens de entrada (origem e destino (I_s, I_t)) são passadas na rede de alinhamento usada para estimar a transformação geométrica G. Em seguida, a contagem *soft-inlier* é calculada (em verde), primeiro encontrando a região de *inliers* (m) de acordo com G e, seguidamente, somando as pontuações de correspondência pareadas dentro dessa área. Rocco et al. (2018) mostraram que a contagem *soft-inlier* é diferenciável, o que permite que todo o modelo seja treinado usando o algoritmo de propagação reversa (*Backpropagation*).

Finalmente, os autores avaliaram sua abordagem em vários bancos de dados de referência para tarefas de correspondência semântica e mostraram que o método proposto –que usa características extraídas com o modelo de classificação ResNet –alcança um desempenho que constitui o estado da arte atual em tarefas de alinhamento semântico até o momento de desenvolvimento desta pesquisa.

2.4 Discussão

Neste capítulo foi apresentado um resumo da evolução das abordagens existentes para a descrição de características relevantes da imagem. Na análise desenvolvida foi evidenciado como os descritores de características baseados em CNN ganharam em desempenho às abordagens artesanais. Também foi mostrado como o desempenho da tarefa de descrição semântica de características da imagem ainda não está à par com a tarefa de classificação de imagens e – geralmente– não existe um cruzamento mutuo entre os modelos construídos para as tarefas de classificação e de descrição de imagens. Ou seja, as abordagens de descritores-CNNs (modelos de descrição baseado em CNNs) apresentadas constroem seus próprios modelos para a extração e a descrição de características e, geralmente, não aproveitam a qualidade (e a elevada acurácia) das características aprendidas pelos modelos CNN de classificação.

Também foram apresentados vários descritores-CNN que aprendem representações de características para vincular de forma robusta (ou fazer correspondência) os pontos semanticamente semelhantes entre diferentes instâncias dentro da mesma categoria de objeto. Esses descritores são normalmente nomeados *descritores semânticos*. A semântica nesse contexto está relacionada com a habilidade que possuem esses modelos de descrição para realizar a correspondência semântica entre as representações aprendidas que são semanticamente semelhantes. A codificação das representações desses modelos é aprendida em banco de dados rotulados para tarefas de correspondência, pelo que não é possível interpretar (pela natureza da aprendizagem) o conteúdo semântico dessas representações.

Ou seja, a interpretação das representações aprendidas por esses descritores semânticos somente tem sentido quando são usadas na tarefa de correspondência semântica, mas as representações por si sós, não são semanticamente interpretáveis. Em consequência, a partir da informação codificada nessas representações, não é possível decodificar essas representações (ou atribuir significado) com o intuito de interpretar alguma informação relacionada à semântica do objeto.

A revisão da literatura mostrou que os modelos CNN de descrição semântica de características representam a informação relevante da imagem usando uma variedade de abordagens diferentes. No entanto, nenhum desses modelos constrói as representações das características do objeto codificando a informação visual da imagem a partir da extensa base teórica da Ciência Cognitiva para representar *o significado*.

Capítulo 3

Teoria dos Protótipos

A Ciência Cognitiva é um paradigma científico contemporâneo que tenta conjugar uma série de campos existentes (inteligência artificial, psicologia, neuro-ciência, filosofia, linguística e antropologia) em um esforço conjunto para estudar o complexo domínio da cognição/inteligência em seu sentido mais amplo; incluindo, por exemplo, problemas de representação do conhecimento, processamento da linguagem, aprendizagem, raciocínio e resolução de problemas, interpretação e representação do significado (Adriaens, 1993).

O presente capítulo apresenta apenas um pequeno resumo de uma das diferentes teorias que propõe a Ciência Cognitiva sobre como os seres humanos representam e relacionam os significados: a Teoria dos Protótipos. Especificamente, se apresentam os aspetos fundamentais que sustentam essa teoria de representação semântica das categorias de objetos, cujos conceitos teóricos tentam ser modelados e reproduzidos nesta pesquisa.

3.1 Semântica

A Semântica (do grego: $\sigma\eta\mu\alpha\nu tik\acute{o}\varsigma$, *semantikós*, "significant") significa o *significado* e a *interpretação* das palavras, da estrutura das sentenças, dos sinais, dos signos, dos objetos, a compreensão e interpretação do mundo, como entendemos os outros e a nós mesmos, e até mesmo quais decisões tomamos como resultado de nossas interpretações (Cuenca & Hilferty, 1999).

Com a revolução científica ocorrida no século XIX em diversas áreas como a biologia, a química, a física e a filologia, inicia-se o estudo do *significado* sob a perspectiva científica. O interesse pela *Semântica* como nova ciência residiu na possibilidade e na necessidade de analisar a representação e interpretação dos significados. A evolução da semântica como ciência foi motivada pela hipótese de que os significados não são cons-

truídos de maneira aleatória e que é necessário entender as leis que os governam (Jaén, 2006).

A semântica evoluiu sob o enfoque de diversas correntes durante o século XX, mas foi na linguística onde se desenvolveram os primeiros avanços com os estudos relacionados a como são atribuídos os significados às palavras. Por exemplo, o estruturalismo como método científico desenvolveu uma corrente semântica (a semântica estrutural) que estabelece que os significados das palavras deve ser estudado respeitando a sua posição com relação ao resto das palavras do sistema (Geeraerts, 1993).

Um trabalho relevante foi a hipótese da *relatividade linguística* ou hipótese Sapir-Whorf (Hoijer & Fearing, 1954; O'Neill, 2015). A hipótese Sapir-Whorf sustenta que a estrutura de uma linguagem afeta a visão do mundo ou a cognição de seus falantes. Essa hipótese foi relevante porque gerou diferentes pesquisas nos anos seguintes que, tentando a sua validação, concluíram outros resultados relevantes. Destacaram-se os estudos de 1969 de Berlin & Kay (1991) sobre a categorização das cores que, precisamente, tentava validar a hipótese Sapir-Whorf. O trabalho de Berlin & Kay (1991) construiu as bases para que a psicóloga americana Eleanor Rosch e a sua equipe (Heider, 1972; Rosch, 1973a, 1975b; Rosch & Mervis, 1975; Rosch et al., 1976; Rosch & Lloyd, 1978) trasladaram esses resultados ao âmbito da psicologia e chegaram a conclusões paralelas, formalizando os conceitos básicos que constituem a Teoria dos Protótipos.

3.2 Categorização como princípio de interpretação

A maioria dos estudos realizados sobre como são construídos os significados pelo ser humano, concluem que são consequência da experiência e do pensamento construído do mundo ao longo de suas vidas (Cuenca & Hilferty, 1999). A atribuição de significados foi analisada sob a perspectiva linguística e uma das características observadas é que quando se atribui um mesmo significado a palavras diferentes (sinônimas), essas palavras são colocadas na mesma categoria cognitiva (Cuenca & Hilferty, 1999). Consequentemente, a compressão ou interpretação da realidade é possível a partir de um conjunto de operações cognitivas complexas, e ao mesmo tempo elementais, que são denominadas *categorização* (Cuenca & Hilferty, 1999).

A categorização é um mecanismo de organização da informação obtida a partir da apreensão da realidade, que é, em si mesma, variada e multiforme (Geeraerts, 1993). A categorização permite simplificar a infinidade do real a partir de dois procedimentos elementares e complementares: a generalização ou abstração, e a discriminação. *Generalizar* é ignorar as diferenças entre entidades e agrupá-las de acordo com suas

semelhanças; enquanto *discriminar* é exatamente o procedimento oposto: insistir nas características diferenciais de duas ou mais entidades para não confundir-las entre si.

Por meio da categorização os seres humanos agrupam diferentes elementos em conjuntos, o que permite pensar, perceber, agir e até mesmo falar (Ungerer & Schmid, 2013). Assim, a categorização pode ser definida como um processo mental de classificação cujos produtos são as *categorias cognitivas*, "conceitos mentais armazenados no cérebro", que, em conjunto e uma vez convencionalizados, "constituem o que se denomina de léxico mental" (Ungerer & Schmid, 2013). Nesse sentido, a categorização fundamenta os processos de interpretação, compreensão e produção linguística (Cuenca & Hilferty, 1999). No entanto, a questão fundamental da maioria das pesquisas não é tanto o que significa *categorizar*, senão "como" esse processo mental inconsciente é realizado e "qual" é a estrutura interna das categorias cognitivas resultantes.

A teoria cognitiva da categorização começa com os trabalhos realizados, principalmente, no campo da antropologia e da psicologia, concretamente em experimentos realizados com as cores. Esses trabalhos foram motivados, geralmente, porque: *i)* as categorias formadas pelas cores são categorias universais de natureza difusa (sem limites definidos entre cada uma delas) e estão codificadas de forma diferente para cada uma das línguas; *ii)* as cores não podem ser definidas com precisão (onde termina o azul e começa o verde?; o que é turquesa: azul ou verde?).

O estudo antropológico de Berlin & Kay (1991) sobre as cores constitui um dos trabalhos mais relevantes. Esses autores realizaram um experimento com falantes de 20 línguas diferentes que deviam identificar (classificar) as cores às quais referem-se os nomes das cores básicas de cada uma das línguas. Berlin & Kay (1991) destacaram como, apesar da grande diversidade de nomes de cores, o número de termos básicos de cores é realmente limitado em cada idioma, e nenhuma linguagem contém mais de 11 nomes básicos de cores. Esses autores chamaram aos pontos no espaço de cor onde acontecem os votos mais generalizados como "pontos focais". Berlin & Kay (1991) concluíram que a categorização linguística das cores não é arbitrária nem é determinada pelas palavras referentes a cada tonalidade em um idioma específico, senão que é baseada nas cores focais (aquelas básicas e mais claramente diferenciadas).

Eleanor Rosch e a sua equipe (Heider, 1972; Rosch, 1973a, 1975b; Rosch & Mervis, 1975; Rosch et al., 1976; Rosch & Lloyd, 1978) levaram esses resultados antropológicos para o campo da psicologia e chegaram a conclusões paralelas sobre a centralidade e a importância da percepção dos "focos" cromáticos ou cores focais. Em um nível teórico elementar, concluíram que existem áreas do espaço da cor que são, no que se refere a percepção, mais "relevantes" do que outras, e que essas áreas são melhor codificadas linguisticamente, pelo que podem ser lembradas mais facilmente (Rosch, 1973a, 1977,

1988). Outro grupo de experimentos permitiu verificar empiricamente a existência de exemplos *bons* e *maus* de elementos da mesma categoria. Esse resultado, diferente do que é inferido na concepção tradicional, mostrou que nem todos os membros de uma categoria têm o mesmo *status*, e que a categoria não pode ser definida com base em condições necessárias e suficientes, comuns a todos os seus membros. Os resultados dos experimentos de Rosch formalizaram muitos dos conceitos que constituem a Teoria dos Protótipos.

3.3 A Teoria dos Protótipos

Umas das contribuições mais relevantes da semântica cognitiva no estudo dos significados das palavras é a Teoria dos Protótipos, a qual visa explicar o mecanismo de categorização e a estrutura interna das categorias. Mesmo quando essa concepção foi analisada inicialmente sob a perspectiva filosófica nos estudos de Wittgenstein (1953), a concepção da categorização baseada em protótipos e a análise da estrutura interna das categorias originou-se no meados da década de 1970 com as pesquisas psicolinguísticas de Eleanor Rosch e um grupo de colegas pesquisadores (Rosch, 1973a,b; Rosch & Mervis, 1975; Rosch, 1975a,b; Rosch et al., 1976; Rosch, 1977; Rosch & Lloyd, 1978; Mervis & Rosch, 1981; Rosch, 1988).

Devido às origens psicolinguísticas, a Teoria dos Protótipos moveu-se em duas direções. Na primeira, a Teoria dos Protótipos teve um sucesso cada vez maior com o desenvolvimento da Linguística Cognitiva desde meados da década de 1980 (Lakoff & Kövecses, 1987; Geeraerts, 1989). Na segunda direção, as descobertas e propostas de Rosch foram abordadas pela psico-lexicologia formal (e mais geralmente, pela psicologia experimental), e vários são os trabalhos que tentaram elaborar modelos formais para a memória conceitual humana e o seu funcionamento (Medin & Schaffer, 1978; Smith & Medin, 1981; Medin & Smith, 1984; Estes, 1986; Minda & Smith, 2002; Zaki et al., 2003). Essas abordagens se aproximaram à Inteligência Artificial.

A Teoria dos Protótipos propõe um modelo de categorização e representação semântica das categorias alternativo ao modelo tradicional baseado na lógica de Aristóteles (Barnes et al., 1995). Ou seja, contrário à crença de que as categorias são *homogêneas* (onde os elementos e as características possuem a mesma relevância na categoria) e *discretas* (que existe um conjunto numerável e bem delimitado de categorias e características do objeto), a Teoria dos Protótipos propõe uma concepção das categorias como *heterogêneas* (nas quais existem alguns membros ou características mais representativos que outros dentro da categoria) e *não discretas*.

Por exemplo, contrário da modelagem que define (ou representa) o significado da categoria *ave* como aqueles elementos com as características penas, pico, capacidade de voar, etc; a Teoria dos Protótipos considera que não existem definições necessárias e suficientes para definir uma categoria específica (Rosch, 1973a, 1977, 1988; Geeraerts, 2010), e considera à categoria como elementos diferentes que têm um estado desigual de representação semântica (ou tipicidade). Por exemplo, sob essa perspectiva, pode-se entender que: *a)* o pardal é mais prototípico da categoria *ave* que o pinguim; *b)* o *Golden Retriever* é mais representativo da categoria *cão* que o chihuahua; e *c)* o *goldfish* é mais típico da categoria *peixe* que o tubarão.

3.3.1 O protótipo e a estrutura interna da categoria

O *protótipo* foi inicialmente definido por Rosch (Heider, 1972; Rosch, 1973b) para denotar um estímulo que assume uma posição saliente na formação de uma categoria; o *protótipo* é o primeiro estímulo associado a essa categoria. Posteriormente, Rosch (1978) definiu o *protótipo* como o elemento que melhor reconhece-se, o mais representativo e distintivo de uma categoria, uma vez que é o elemento que compartilha mais características com os outros membros da categoria e menos com os membros de outras categorias.

Rosch (Rosch, 1978; Rosch & Lloyd, 1978) concluiu que à tendência de definir categorias de forma rígida, entra em contradição com a percepção psicológica. O conceito prototípico de Rosch (Rosch, 1978; Rosch & Lloyd, 1978) mostrou que as categorias baseadas em percepção não possuem limites bem delimitados. Os experimentos de Rosch mostraram que, contrário à concepção das categorias como partições fixas claramente delimitadas e definidas, as categorias cognitivas são entidades difusas, onde o passo de uma categoria para outra é gradual e é marcado pelos *membros periféricos*.

A abordagem baseada em protótipos de Rosch sobre a estrutura semântica das categorias concentra-se nos tipos de fenômenos que foram observados por Erdmann (1910 apud Geeraerts (2010)) ou Gipper (1959 apud Geeraerts (2010)), mas que praticamente não receberam atenção teórica sistemática naquele contexto: "as categorias linguísticas podem ser difusas nas bordas, mas claras no centro" (Geeraerts (2010)). Um experimento realizado por Gipper (1959 apud Geeraerts (2010)) para estudar os significados das palavras alemãs *Stuhl* (*cadeira*) e *Sessel* (*cadeira confortável*) mostra que, dentro de uma categoria, o significado semântico central pode mudar de acordo com a importância da característica observada (Figura 3.1).

Os experimentos de Rosch (Rosch, 1973a, 1977, 1978; Rosch & Lloyd, 1978) mostraram que em vez de demarcações claras entre áreas conceituais igualmente importan-

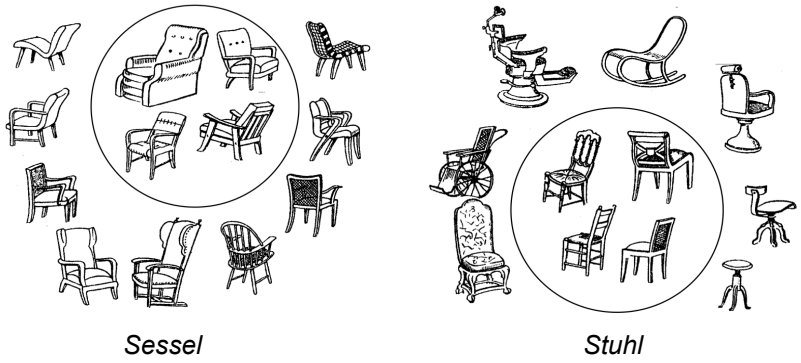


Figura 3.1: O experimento *Sessel* e *Stuhl* conduzido por Gipper. O experimento estuda os significados das palavras alemãs *Stuhl* (cadeira) e *Sessel* (cadeira confortável). Mostra-se que, dentro da categoria *cadeira*, o conceito semântico central pode mudar dependendo da importância da característica observada. Esse fenômeno é descrito na semântica psicolinguística como *organização prototípica* e constitui uma das motivações principais desta pesquisa. Fonte: Adaptado de Geeraerts, 2010, p. 297.

tes, encontraram-se áreas marginais entre categorias que somente são definidas inequivocamente em seus pontos focais. Consequentemente, Rosch (Rosch, 1978) definiu às categorias como grupos de objetos do mundo relacionados por causa das semelhanças que mantêm uns com outros, organizados em torno de uma imagem central, prototípica, do membro da categoria que é o mais representativo de todos. O pertencimento de um elemento a uma categoria é estabelecido a partir do grau de similaridade com o protótipo, embora os atributos comuns entre o elemento em questão e o protótipo não devem ser entendidos como condições necessárias e suficientes de toda a categoria. Rosch desenvolveu essa observação em uma visão prototípica mais geral das categorias de linguagem natural, particularmente, para *categorias que denominam objetos naturais* (Rosch et al., 1976; Mervis & Rosch, 1981; Geeraerts, 2010).

A definição do protótipo proposta por Rosch gerou avanços no entendimento do mecanismo de categorização e na representação da estrutura interna das categorias, mas mostrou ser insuficiente quando foi aplicada à grande diversidade de categorias existentes (Cuenca & Hilferty, 1999). Cada categoria deve ter um e apenas um protótipo? O protótipo tem uma ou mais características em comum com os outros membros da categoria? Essas são apenas algumas perguntas que geraram a introdução de conceitos como *semelhança familiar* e *efeitos prototípicos*.

3.3.2 Semelhança familiar

Rosch & Mervis (1975) adotaram o conceito de semelhança familiar (*family resem-*

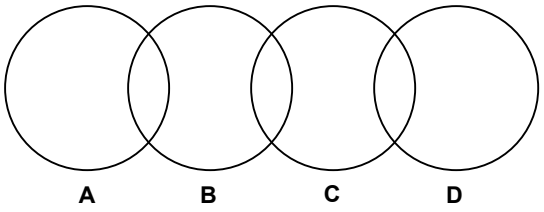


Figura 3.2: Modelo de semelhança familiar de Wittgenstein. Fonte: Cuenca & Hilferty, 1999, p. 38.

blance) de Wittgenstein (1953) para complementar os conceitos da Teoria dos Protótipos formulados anteriormente. As autoras mostraram que o conceito de semelhança familiar constitui um dos principais princípios estruturais que regem a formação da estrutura prototípica das categorias semânticas.

Para Wittgenstein (1953 apud Cuenca & Hilferty (1999)) as categorias não são discretas e absolutas, conforme estabelecido pela tradição filosófica aristotélica, senão que são difusas e heterogêneas. O conceito de *semelhança familiar* sustenta que os elementos da categoria que poderiam estar conectados por uma característica comum essencial, podem, de fato, estar conectados por uma série de similaridades sobrepostas, onde nenhuma característica é comum a todos os elementos.

Wittgenstein (1953 apud Cuenca & Hilferty (1999)) argumentou que os referentes de uma palavra não precisam ter elementos comuns para serem compreendidos e utilizados no funcionamento normal da linguagem. O autor sugeriu que, pelo contrário, uma semelhança familiar pode ser a que conecta os vários referentes de uma palavra. Uma relação de *semelhança familiar* assume a forma AB, BC, CD (Rosch & Mervis, 1975). Ou seja, cada elemento tem pelo menos uma, e provavelmente várias, características em comum com um ou mais elementos; mas nenhuma, ou poucas características, são comuns a todos os elementos (Geeraerts, 2010)(Ver Figura 3.2).

Rosch & Mervis (1975) contextualizaram o conceito de *semelhança familiar* na Teoria dos Protótipos mostrando que as associações entre os membros de uma categoria não são necessariamente estabelecidas entre as entidades da categoria e o protótipo. As autoras mostraram que é possível que um elemento seja integrado à categoria por sua semelhança com outro membro que sim possui algum atributo comum com a imagem mental do protótipo. Ou seja, não é necessário que todos os membros de uma categoria tenham algum atributo comum entre si, nem algum atributo comum com o protótipo, senão que as possibilidades associativas sejam múltiplas (Cuenca & Hilferty, 1999).

Por exemplo, sejam $\{a, b, c, d\}$ atributos representativos da categoria C. A avaliação qualitativa da relevância de cada atributo para a categoria pode definir-se como:

	a	b	c	d
elemento 1 (prototípico)	+	+	+	+
elemento 2	+	+	+	-
elemento 3	+	+	-	+
elemento 4	+	-	-/+	+

Tabela 3.1: *Caracterização dos elementos da categoria C.* A caracterização de cada membro da categoria foi realizada com o signo "+" (se possui o atributo) e com o signo "-" (se não possui o atributo). O elemento prototípico da categoria possui todos os atributos relevantes da categoria. Fonte: Adaptado de Geeraerts, 2010, p. 305.

<b> é um atributo característico da categoria; <a> é um atributo geral que pode aparecer em outras categorias, <d> é um atributo relevante da categoria (mas não característico) e <c> é um atributo que pertence ao modelo idealizado da categoria C. Sejam os elementos {*elemento 1, elemento 2, elemento 3, e elemento 4*} membros da categoria C caracterizados com os atributos que pertencem à categoria como se apresenta na Tabela 3.1.

Assim, de acordo com a Teoria dos Protótipos, as categorias semânticas são formadas pela interseção de uma ou várias propriedades típicas, que tendem coincidir, embora tal coincidência não seja estritamente necessária (Ver exemplo de construção na Figura 3.3 para o caso da categoria C mostrada na Tabela 3.1).

Na Figura 3.3, a zona central representa os membros que possuem as quatro características que definem o protótipo da categoria C. A partir desse núcleo, o *continuum* categorial pode ser caracterizado por duas gradações: *i)* cada característica é avaliada pela importância relativa que possui para a categoria; e *ii)* a relevância (ou saliência) de cada membro da categoria está de acordo com a quantidade e tipo de característica que o elemento apresenta. Dessa forma, é possível estabelecer o *grau de prototipicidade* de um determinado elemento dentro da categoria (Rosch & Mervis, 1975).

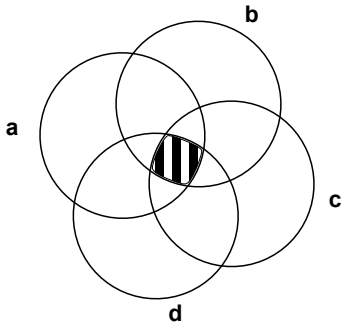


Figura 3.3: Modelo de semelhança familiar da Teoria dos Protótipos. Fonte: Cuenca & Hilferty, 1999, p. 40.

3.3.3 Efeitos prototípicos

Os resultados do protótipo de Rosch (Heider, 1972; Rosch, 1973b, 1977, 1978) mostraram que era importante distinguir claramente entre os vários fenômenos que podem estar associados à *prototipicidade*. Um dos problemas da definição do *protótipo* como um "protótipo - objeto", ou um elemento que é *prototípico* da categoria, está relacionado à decisão de: quem é o protótipo da categoria quando dois elementos são igual de representativos?

Em consequência, começou a definir-se o protótipo como "protótipo - entidade cognitiva" ou, mais especificamente, como *efeitos de prototipicidade* (Rosch, 1988; Cuenca & Hilferty, 1999; Geeraerts, 2010). Para cada categoria é construída uma imagem mental que pode corresponder –de maneira mais ou menos exata– a um membro existente da categoria, com mais de um membro, ou com nenhum membro em particular. Essa imagem mental ou *representação cognitiva* é o que se denomina *protótipo da categoria* (Mervis & Rosch, 1981; Rosch, 1988; Geeraerts, 1997). Assim, quando o protótipo é referenciado, está especificando-se uma abstração que realmente refere-se aos julgamentos sobre o *grau de prototipicidade* dos elementos da categoria. O *protótipo* é um fenômeno de superfície que toma formas diferentes de acordo com a categoria de estudo; é, basicamente, o produto das representações mentais do mundo, e dos modelos cognitivos idealizados construídos com a experiência (Cuenca & Hilferty, 1999).

Os efeitos prototípicos surgem precisamente de inter-relações imperfeitas entre a realidade e o modelo cognitivo idealizado (Cuenca & Hilferty, 1999). Segundo Rosch (Rosch, 1978; Mervis & Rosch, 1981; Rosch, 1988), quatro características são frequentemente mencionadas como típicas da *prototipicidade* nas categorias: *i)* as categorias prototípicas exibem graus de tipicidade, não todos os membros são igualmente representativos da categoria; *ii)* as categorias prototípicas exibem uma estrutura de semelhança familiar (*family resemblance structure*) ou, mais geralmente, sua estrutura semântica assume a forma de um conjunto radial de leituras agrupadas e sobrepostas (Ver Figura 3.3); *iii)* as categorias prototípicas são difusas (*blurred*) nas bordas; e *iv)* as categorias prototípicas não podem ser definidas por meio de um único conjunto de critérios, ou atributos necessários e suficientes.

Geeraerts (2010) expressa que existe atualmente um consenso na literatura linguística sobre a *prototipicidade* nas categorias prototípicas. Geeraerts (2010) expõe que as quatro características definidas por Rosch como típicas da *prototipicidade* (enumeradas anteriormente) não são necessariamente coextensivas a toda a categoria; e que essas características nem sempre co-ocorrem. Essa nova perspectiva da *prototipicidade* expressa que as características da prototipicidade são *efeitos da prototipicidade* (ou

	extensional (no nível dos exemplares)	*intensional* (no nível de definição)
non-equality (efeitos salientes, núcleo / periferia)	a) diferenças de tipicidade e saliência na categoria	b) agrupamento em semelhanças familiares
non-discreteness (problemas de demarcação, flexibilidade)	c) difusas nas bordas, incerteza de pertencimento à categoria	d) ausência de definições necessárias e suficientes

Tabela 3.2: *Os Efeitos Prototípicos da Teoria dos Protótipos.* a) *extensional non-equality* das estruturas semânticas: alguns membros de uma categoria são *mais típicos* ou mais representativos e mais salientes da categoria do que outros; b) *intensional non-equality*: as leituras de um item léxico podem formar um conjunto com um ou mais casos relevantes envolvidos por leituras periféricas que emanam das leituras centrais e mais salientes; c) *extensional non-discreteness*: podem existir variabilidades na fronteira de uma categoria; d) *intensional non-discreteness*: a demarcação de definições das categorias lexicais pode ser problemática, medido no contexto do requisito clássico de que as definições assumem a forma de um conjunto de atributos necessários que são conjuntamente suficientes para delimitar a categoria. Fonte: Traduzido de Geeraerts, 2010, p. 304.

efeitos prototípicos) que podem ser exibidos em várias combinações por itens lexicais individuais e podem ter fontes muito diferentes (Geeraerts, 2010). Geeraerts (2010) propõe que as quatro características da prototipicidade são sistematicamente relacionadas ao longo de duas dimensões, o que permite redefinir os *efeitos prototípicos* como as quatro características[1]: *i)* não-igualdade extensível (*extensional non-equality*); *ii)* não-igualdade intencional (*intensional non-equality*); *iii)* não-discrição extensível (*extensional non-discreteness*) e *iv)* não-discrição intencional (*intensional non-discreteness*) (Ver Tabela 3.2). O conceito de prototipicidade constitui, por si só, uma agrupação prototípica para essas quatro características em que os conceitos de *non-discreteness* e *non-equality*, tanto no nível *intensional* como *extensional*, desempenham um papel distintivo importante (Geeraerts, 2010).

Rosch (Rosch et al., 1976; Rosch, 1977; Mervis & Rosch, 1981) mostrou que a saliência psicológica de determinadas características perceptuais pode ser estendida a outros domínios, como por exemplo, o domínio dos objetos naturais. Essas descobertas em conjunto com os conceitos de semelhança familiar e os efeitos prototípicos condicionaram que, posteriormente, Rosch (Rosch, 1978; Mervis & Rosch, 1981; Rosch, 1988) e Geeraerts (2010) definiram formalmente que os casos centrais claros de uma categoria constituem o *protótipo* da categoria, mas a categoria não precisa estar tão delimitada

[1]Doravante usam-se os termos em inglês para evitar ambiguidades conceituais.

como esse centro semântico. Assim, a definição semântica (ou alcance da aplicação) de tais categorias concentra-se em pontos focais redondos representados por membros prototípicos da categoria. Os atributos desses membros focais são as propriedades estruturalmente mais salientes da categoria em questão e, inversamente, um membro específico da categoria ocupa uma posição focal porque exibe as características mais salientes da categoria.

Os *efeitos prototípicos* conjeturam a importância da distinção entre *significado central* e *significado periférico*. Nessa perspectiva, entende-se que a presença do significado prototípico central é quem regula que os significados (e as características que os definem) não apareçam de um modo totalmente arbitrário (Rosch, 1988; Geeraerts, 2010). Ou seja, sempre deve existir algum contato (seja direto ou indireto) entre os novos significados e o protótipo, estabelecendo uma semelhança familiar entre os distintos significados. Os efeitos prototípicos fundamentam as mudanças semânticas e justificam como os membros de uma categoria são distribuídos dentro da mesma (Ver exemplo no Anexo B). Sobre esses aspectos teóricos, a Semântica Cognitiva constitui um modelo baseado em protótipos que torna sistemático e lógico aquilo que era inconsistente ou mesmo inexplicável para as teorias semânticas anteriores (Jaén, 2006).

3.4 Modelo do Protótipo da Psicologia Experimental

As origens psicolinguísticas da Teoria dos Protótipos geraram que muitos dos conceitos introduzidos por Rosch (Rosch et al., 1976; Rosch, 1977; Mervis & Rosch, 1981) foram herdados pela Psicologia Experimental para construir uma linha de pesquisa com uma perspectiva diferente da Teoria dos Protótipos da Linguística Cognitiva (Lakoff & Kövecses, 1987; Geeraerts, 1989). Os principais trabalhos dessa nova linha de pesquisa aproximaram-se muito à Inteligência Artificial e propuseram abordagens que tentaram elaborar modelos formais para simular e explicar a memória conceitual humana e o seu funcionamento (Posner & Keele, 1968; Reed, 1972; Homa & Vosburgh, 1976; Medin & Schaffer, 1978; Smith & Medin, 1981; Medin & Smith, 1984; Estes, 1986; Nosofsky, 1986; Shepard, 1987; Minda & Smith, 2001, 2002; Zaki et al., 2003).

Os principais trabalhos apresentados nessa linha de pesquisa propuseram modelos que tentavam representar as categorias e usavam essas representações na categorização. Nesse sentido, duas abordagens têm tentado responder como é realizada –pelos seres humanos– a representação da categoria no processo de categorização: *i)* em termos de protótipos abstratos (*prototype model*); ou *ii)* em termos de exemplos específicos

da categoria (*exemplar model*). De acordo com o modelo do protótipo (*prototype model*) (Reed, 1972; Homa & Vosburgh, 1976; Minda & Smith, 2001), as pessoas representam categorias em termos de alguma tendência central calculada sobre as instâncias de treinamento da categoria e classificam os objetos com base em quão semelhantes são aos protótipos das categorias aprendidas. Em contraste, de acordo com o modelo dos exemplares (*exemplar model*) (Medin & Schaffer, 1978; Nosofsky, 1986; Zaki et al., 2003), as pessoas representam categorias armazenando as próprias instâncias de treinamento individuais e classificam os objetos baseado na similaridade com cada um desses exemplares.

Um paradigma experimental que tem sido extremadamente usado para contrastar as predições do modelo dos exemplares e do modelo do protótipo é a estrutura de categoria 5/4 (*5/4 category structure*) de Medin & Schaffer (1978). Nesse paradigma, foram coletados um conjunto de estímulos psicológicos constituídos por valores binários em cada uma das quatro dimensões: cor [vermelho ou azul], forma [triângulo ou círculo], tamanho [grande ou pequeno] e número de elementos [um ou dois]. Assim, os estímulos usados pelos autores são formas perceptivas simples que variam ao longo de quatro dimensões salientes de valor binário. O conjunto de estímulos foram classificados em duas categorias (Categorias A e B) e divididos em dados de teste e treinamento. Os valores lógicos do protótipo da Categoria A são assumidos como 0 0 0 0 (um triangulo grande de cor vermelho), e os valores lógicos do protótipo da Categoria B são assumidos como 1 1 1 1 (dois círculos pequenos de cor azul).

A notação usada e os valores binários de cada característica dos estímulos geraram que a estrutura de categoria 5/4 fosse o banco de dados padrão usado para o diagnóstico e comparação das famílias de modelos do protótipo e dos exemplares (Minda & Smith, 2002; Zaki et al., 2003). Apesar da grande variedade de modelos existentes para cada família (modelo do protótipo e modelo dos exemplares), nesta seção será apresentada uma breve revisão dos modelos formais principais que sustentam a presente pesquisa.

3.4.1 O Modelo de Contexto Generalizado

De acordo com o Modelo de Contexto Generalizado (*Generalized Context Model (GCM)*) (Nosofsky, 1986), as pessoas representam categorias armazenando exemplares individuais na memória. Os exemplares são representados como pontos (ou estímulos) em um espaço multidimensional. Como os estímulos na estrutura de categoria 5/4 variaram ao longo de quatro dimensões separáveis de valor binário, o autor propôs que a *distância psicológica* entre dois estímulos i, j pode ser calculada como a distância

ponderada:

$$d_{ij} = \sum_{m=1}^{4} \omega_m \left| x_{im} - x_{jm} \right|, \tag{3.1}$$

onde: $x_{im} \in [0, 1]$ é o valor do estímulo na m-dimensão, e ω_m ($0 \leq \omega_m \leq 1, \sum \omega_m = 1$) é o custo (ou peso) de atenção dado à dimensão m. Consequentemente, a similaridade entre dois estímulos i, j é uma função de decaimento exponencial de sua distância psicológica:

$$S_{ij} = exp(-\alpha d_{ij}), \tag{3.2}$$

onde α é um parâmetro de *sensibilidade* (Shepard, 1987). Nosofsky (1986) propôs que a probabilidade de que o estímulo i seja classificado na Categoria A é dada pela expressão:

$$P(A|i) = \frac{(\sum S_{ia})^{\gamma}}{(\sum S_{ia})^{\gamma} + (\sum S_{ib})^{\gamma}}, \tag{3.3}$$

onde γ constitui um parâmetro de escala de resposta (*response-scaling parameter*), e $\sum S_{ia}$ e $\sum S_{ib}$ são as somas das similaridades do item i com todos os exemplares das Categorias A e B, respectivamente.

3.4.2 O Modelo do Protótipo Multiplicativo

O Modelo do Protótipo Multiplicativo (*Multiplicative Prototype Model (MPM)*) (Minda & Smith, 2001, 2002) é uma versão do *modelo do protótipo* original (Reed, 1972; Homa & Vosburgh, 1976), mas que usa as mesmas funções de similaridade do modelo GCM. Nesse modelo, a distância semântica entre o item i e o protótipo da Categoria A é dado por:

$$d_{iA} = \sum_{m=1}^{4} \omega_m \left| x_{im} - P_{Am} \right|, \tag{3.4}$$

onde P_{Am} denota o valor do Protótipo A na dimensão m, e os valores ω_m são os pesos da atenção dada à dimensão m. A semelhança entre o item i e o protótipo da Categoria A é dada pela expressão: $S_{iA} = exp(-\alpha d_{iA})$, onde α é o mesmo parâmetro de *sensibilidade* da Equação 3.2. No modelo MPM a probabilidade com que o item i é

classificado na Categoria A é dada pela equação:

$$P(A|i) = \frac{S_{iA}^{\gamma}}{S_{iA}^{\gamma} + S_{iB}^{\gamma}}, \tag{3.5}$$

Doravante, nesta pesquisa, é proposta uma metodologia para representar computacionalmente os protótipos semânticos das categorias de objetos. A metodologia proposta generaliza vários dos conceitos dos modelos formais GCM e MPM apresentados.

3.5 Usos na Ciência da Computação

As bases teóricas das correntes semânticas mencionadas anteriormente possibilitaram interlocuções interdisciplinares, visando aplicar esses princípios teóricos estruturais e cognitivos em outros âmbitos científicos. Por exemplo, no campo da Ciência da Computação, a Inteligência Artificial baseou-se nos postulados da teoria semântica e a linguística para o desenvolvimento de algumas de suas áreas de pesquisa. Os preceitos semânticos foram usados, por exemplo, em Linguística Computacional contribuindo com o processo de análise e síntese de um texto de forma eficiente e autônoma por um computador (Montague, 1973; Collobert & Weston, 2008; Goodman et al., 2014; Erk, 2016; Boleda & Herbelot, 2016). A área de Processamento da Linguagem Natural (*Natural Language Processing (NLP)*), embora tenha suas origens nos trabalhos de Alan Turing de 1950, desenvolveu-se com mais eficiência, nos anos seguintes, com o surgimento de novas e poderosas ferramentas de processamento (LeCun et al., 2015).

A semântica, no contexto dessas áreas, assume que o significado de uma sentença é equivalente a suas *condições de verdade*. A descrição semântica de uma linguagem estabelece um mecanismo capaz de determinar as condições da verdade para cada sentença. As ferramentas básicas para o desenvolvimento de teorias semânticas nesse âmbito procedem da lógica e das *fórmulas bem formadas*, assim como da representação do conhecimento. As aplicações mais significativas resultantes dessa área do processamento de texto manifestam-se na compreensão da linguagem, na recuperação de informação, na extração de informação, na busca de respostas, na tradução automática, na reconstrução de discurso, no reconhecimento de fala, dentre outras.

Baseado no sucesso da aplicação das técnicas anteriores, a Visão Computacional - outra área de estudo da Inteligência Artificial - herdou e aplicou alguns desses métodos e enfoques para dar respostas a desafios semânticos envolvendo imagens. Na década de 1960, os primeiros pesquisadores dessa área acreditaram que a tarefa de interpretação e

processamento de imagens por computador seria fácil, o que não se confirmou e muitos anos de pesquisa têm demonstrado que continua sendo um desafio muito complexo.

Surgiram nesse período alguns trabalhos incipientes que introduziram abordagens semânticas para analisar imagens. Destacou-se, por exemplo, o estudo de Waltz (1975) que propôs um método de geração de descrições semânticas a partir da construção de cenas com sombras, oferecendo - entre suas contribuições - explicações teóricas aos trabalhos heurísticos da análise da cena que foram propostos principalmente por Guzman-Arenas (1968) e Winston (1970).

Especificamente, a Teoria dos Protótipos não tem sido amplamente utilizada em Visão Computacional, embora existam alguns estudos na literatura de Inteligência Artificial (Malisiewicz & Efros, 2008; Malisiewicz, 2011; Zhao & Qin, 2015; Saleh et al., 2016) que apresentam algumas tentativas de aplicação. Destacam-se os trabalhos (Kuncheva & Bezdek, 1998; Fernández & Isasi, 2004; Kim & Oommen, 2004; Zhang et al., 2012; Wohlhart et al., 2013) que desenvolveram métodos de Aprendizagem de Máquina que introduziram o protótipo em tarefas de classificação, mas esses protótipos aprendidos não eram regidos pelos conceitos principais da Teoria dos Protótipos.

Zhao & Qin (2015) propuseram uma nova métrica de dissimilaridade baseada em um *framework* de representação do conhecimento denominado "rótulo semântico" interpretado pela Teoria dos Protótipos. A métrica de dissimilaridade proposta pelos autores é utilizada para complementar o algoritmo clássico K-means para agrupamento de instâncias de dados e dos conceitos vagos representados por expressões lógicas de rótulos linguísticos.

Saleh et al. (2016) mostraram que os modelos de aprendizagem profunda não podem generalizar imagens atípicas que são substancialmente diferentes das imagens de treinamento. Esses autores concluíram que o envolvimento de informações sobre a tipicidade/atipicidade das amostras de treinamento como um termo de ponderação na função de perda, ajuda muito a melhorar o desempenho em exemplos atípicos "invisíveis", quando o treinamento é feito apenas com exemplos típicos. Saleh et al. (2016) inspiraram-se na Teoria dos Protótipos para mostrar que a aprendizagem com maior ênfase em amostras representativas (típicas) aumenta a capacidade de generalização dos classificadores CNN treinados de maneira discriminativa.

O protótipo, como representação, tem sido utilizado em algumas abordagens de classificação. Trabalhos como os de Crammer et al. (2003); Kohonen (1988); Seo & Obermayer (2003); Wohlhart et al. (2013) representam amostras de dados, atribuindo-os a um conjunto de protótipos. Essa abordagem divide o espaço das características de entrada e tem a vantagem de reduzir a complexidade dos classificadores restringindo a análise aos protótipos selecionados. O desempenho dessa abordagem depende

fortemente da escolha correta dos protótipos (Wohlhart et al., 2013).

Learning Vector Quantization (LVQ) é uma área iniciada pelo trabalho seminal de Kohonen (1988), no qual os métodos tentam encontrar os melhores protótipos a partir de dados rotulados. Esse trabalho foi modificado por Crammer et al. (2003); Seo & Obermayer (2003); Wohlhart et al. (2013), mas a heurística geral – ou método de aprendizagem – tenta mover os protótipos próximos às amostras de treinamento da mesma categoria e distantes das outras categorias. Wohlhart et al. (2013) aprendem os protótipos usando uma formulação softmax de um relaxamento probabilístico da classificação 1-NN. Os autores aprendem os protótipos usando o gradiente descendente de forma discriminativa e, durante o teste, avaliam o pertencimento dos membros às categorias comparando apenas as distâncias aos poucos protótipos já aprendidos.

Trabalhos recentes usam os protótipos para melhorar a classificação das abordagens *zero-shot learning* (Jetley et al., 2015; Snell et al., 2017) e *few-shot learning* (Snell et al., 2017). Jetley et al. (2015) propõem uma arquitetura de rede que introduz o protótipo como conhecimento *a priori*. Esses autores assumem como protótipo da categoria um conjunto de imagens fixas (*templates*) e cria um espaço integrado comum (*common embedding space*), no qual a similaridade entre cada par <*protótipo, instância de objeto*> é calculada usando o produto escalar.

Snell et al. (2017) propõem a Rede Prototípica (*Prototypical Network*) que aprende um espaço métrico no qual a classificação pode ser realizada calculando as distâncias para as representações dos protótipos de cada categoria. A Rede Prototípica aprende um mapeamento não linear da entrada em um espaço integrado (*embedding space*). Esses autores calculam os protótipos como o vetor médio dos pontos (do espaço integrado) mapeados por meio de uma função com parâmetros aprendidos.

Em geral, os trabalhos de classificação apresentados introduzem a informação prototípica como o centro do processo de aprendizagem dos modelos de classificação. Alguns modelos (Seo & Obermayer, 2003; Wohlhart et al., 2013) aprendem o protótipo, e outros (Snell et al., 2017) calculam os protótipos no espaço integrado aprendido.

3.6 Discussão

O presente capítulo apresenta a evolução dos principais conceitos da Teoria dos Protótipos que tentam justificar a representação do significado das categorias cognitivas. Foram apresentados o conceito de protótipo e as características típicas relacionadas com a prototipicidade: esses conceitos relevantes visam justificar as mudanças semânticas e a estrutura interna das categorias prototípicas.

Aliás, se apresentaram os principais modelos da Psicologia Experimental que visam reproduzir os conceitos principais da Teoria dos Protótipos, e que constituem o baseamento teórico do Modelo Computacional do Protótipo que se propõe nesta pesquisa.

Na última seção do Capítulo analisaram-se os trabalhos mais relevantes da Ciência da Computação que tentaram usar alguns dos conceitos da Teoria dos Protótipos para construir algoritmos de Aprendizagem de Máquina. Alguns trabalhos aprenderam uma métrica de dissimilaridade, outros usaram o conceito de tipicidade de um objeto para aumentar o poder de generalização das CNN, outros aprenderam o protótipo para aumentar a acurácia de modelos de classificação; mas nenhum desses trabalhos usou –em conjunto– todos os conceitos introduzidos pela Teoria dos Protótipos para representar a semântica das categorias aprendidas.

Capítulo 4

Metodologia Geral

A Teoria dos Protótipos propõe um modelo de representação do significado das categorias de objetos onde o protótipo da categoria constitui a entidade semântica que regula a estrutura semântica interna das categorias de objetos. O protótipo da categoria encapsula as características típicas (ou definições) que representam semanticamente à categoria e permite que os membros da categoria sejam distribuídos dentro da categoria baseados na sua tipicidade e agrupados por sua semelhança familiar. Ou seja, o protótipo regula que o significado atribuído a cada objeto (e às características que os definem) dentro da categoria não seja gerado de modo totalmente arbitrário, pois para definir a representatividade do objeto sempre deve existir algum contato (seja direto ou indireto) entre os novos significados e o protótipo da categoria. Consequentemente, de acordo com *o modelo do protótipo* (Posner & Keele, 1968; Reed, 1972; Homa & Vosburgh, 1976; Minda & Smith, 2001; Zaki et al., 2003), os seres humanos detectam e reconhecem objetos com base nas semelhanças/diferenças das características observadas do objeto com relação aos *protótipos* aprendidos. As categorias de objetos não aprendidas (protótipo desconhecido) não são reconhecidas corretamente.

Sob essas premissas da Teoria dos Protótipos, o propósito principal desta pesquisa consiste em introduzir a informação semântica dos *protótipos* das categorias de objetos na descrição semântica global de objetos. Ou seja, se a semântica encapsulada no protótipo da categoria pode reger o processo de classificação semântica de objetos, por que não usar essa representação semântica da categoria (o protótipo) para descrever semanticamente o objeto? Se o protótipo da categoria pode posicionar semanticamente o elemento em uma posição específica (e única) dentro da estrutura semântica interna da categoria, por que não usar o protótipo para encontrar quais são essas características do objeto que geram que seja posicionado nessa posição específica dentro da categoria?

Especificamente, pretende-se seguir uma abordagem de descrição semântica que

simula o comportamento humano nas tarefas de descrição global de objetos. O intuito da abordagem proposta é tentar simular os processos de generalização e descriminação semântica para usar a estratégia humana de descrever globalmente um objeto destacando as características que o tornam único dentro da categoria à qual pertence (Vide Figura 1.2). Observa-se que usando essa abordagem, primeiramente o objeto deve ser reconhecido (entenda-se recuperar a informação semântica do protótipo da categoria à qual o objeto pertence) e somente depois é possível descrever as características do objeto que o tornam distintivo dentro da categoria.

Esta pesquisa propõe abordar o problema da descrição semântica global de objetos através da construção de uma representação semântica das características do objeto que encapsula o *significado semântico* do objeto (a semântica usada para categorizá-lo) e a *distinção semântica* do objeto (as características que distinguem o objeto dentro da categoria). A abordagem que se apresenta assume que é possível construir uma representação semântica do protótipo da categoria. Propõe-se usar essa representação semântica para encontrar a *distinção semântica* do objeto (*diferença semântica* com relação ao protótipo da categoria).

A metodologia apresentada propõe uma perspectiva diferente para abordar o problema da descrição semântica de objetos. Dissimilar às abordagens atuais que utilizam as CNNs para a descrição de características, esta pesquisa propõe um modelo que constrói a descrição global do objeto codificando a informação da imagem baseado nos fundamentos teóricos da Teoria dos Protótipos para representar *o significado*. Especificamente, o método de descrição semântica global proposto se destaca pela:

- Simulação dos conceitos principais da Teoria dos Protótipos: construção do protótipo, representação da estrutura interna das categorias de objetos e categorização de objetos baseada nos protótipos aprendidos (*a priori*).

- Utilização do *protótipo* construído (significado semântico central da categoria) como entidade semântica (ou sequência de DNA) que rege a codificação e representação das assinaturas semânticas dos objetos que pertencem à categoria.

- Utilização das características aprendidas pelos modelos CNN de classificação pré-treinados como as *característica do objeto*, visando descrever semanticamente o objeto usando as mesmas características usadas para classificá-los.

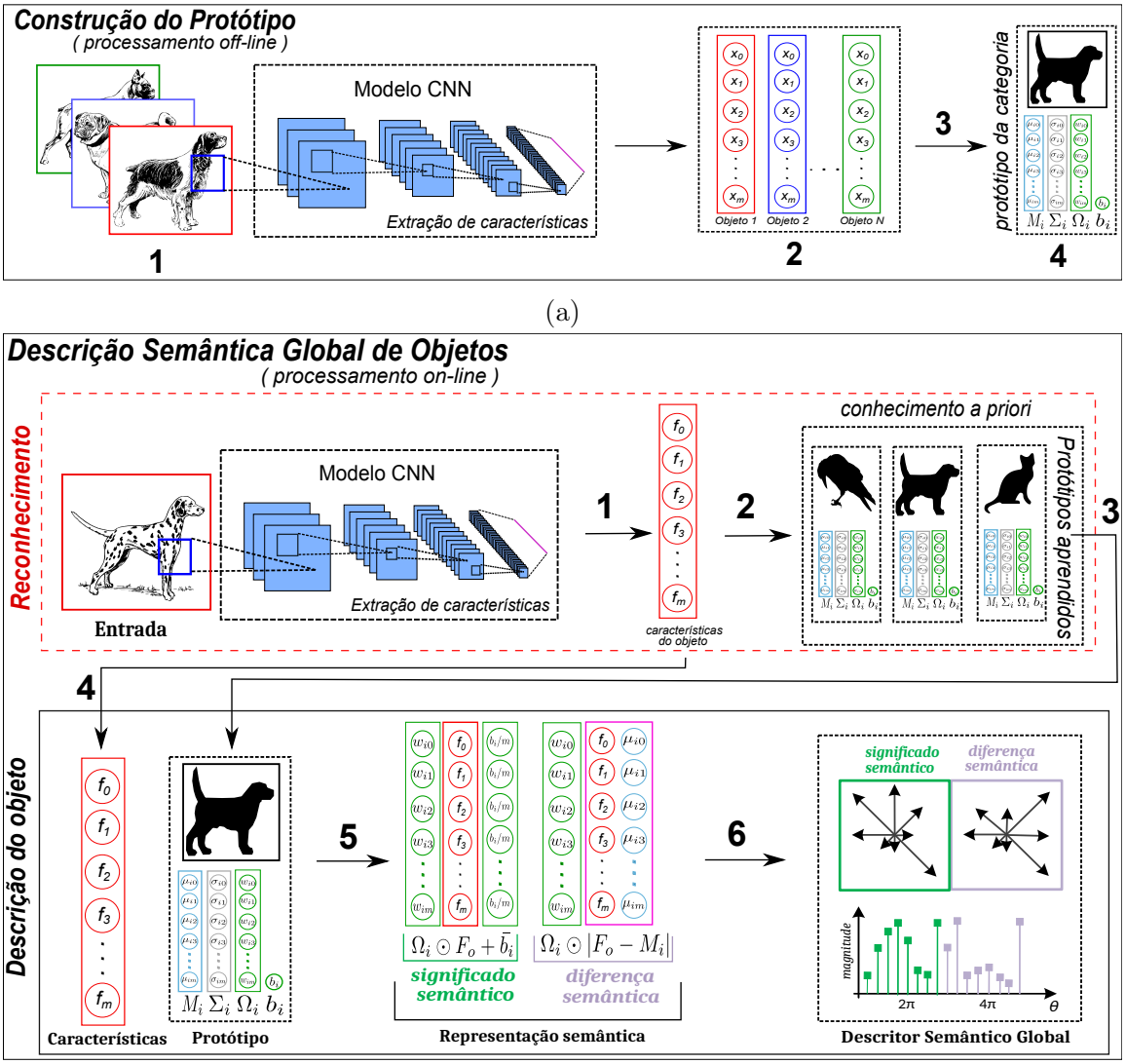


Figura 4.1: *Visão geral da metodologia proposta.* a) *Construção do protótipo da categoria a partir das características extraídas dos modelos CNN de classificação*: *1)* exemplares da categoria cão; *2)* características extraídas de cada elemento da categoria; *3)* construção do protótipo usando as características dos objetos e a relevância (pesos) de cada característica para a categoria; *4)* armazenamento apenas do protótipo da categoria. b) *Modelo de descrição semântica global de objetos baseado em protótipos proposto*: *1)* extração das características do objeto de entrada; *2)-3)* reconhecimento (categorização) do objeto baseado em protótipos; *4)* características do objeto; *5)* construção da *representação semântica do objeto* a partir das características do objeto e do protótipo da categoria; *6)* redução da dimensionalidade da *representação semântica do objeto* e construção da assinatura do Descritor Semântico Global. Fonte: Elaborado pelo autor.

A Figura 4.1 apresenta a visão geral da metodologia proposta, que visa simular o fluxo conceitual da hipótese de descrição semântica global de objetos baseada em

protótipos (Vide Figura 1.2). Nessa sinopse gráfica são exibidos os passos que devem ser desenvolvidos para alcançar o objetivo geral da pesquisa. Observe-se que a metodologia de descrição semântica proposta está constituída por duas etapas principais: *a)* uma primeira etapa onde são preparados e armazenados os protótipos das categorias de imagens de objetos para serem usados conforme seja necessário (processamento *off-line*) (Ver Figura 4.1a); *b)* uma última etapa onde a representação semântica do objeto de entrada é construída usando a abordagem proposta de descrição semântica global de objetos baseada em protótipos (processamento *on-line*) (Ver Figura 4.1b).

Na Figura 4.1a são apresentados os principais passos desenvolvidos para a construção do *protótipo* da categoria a partir do modelo matemático proposto para *codificar* e *armazenar* (representar) o *significado semântico central* da categoria. O protótipo da categoria é construído usando as características extraídas dos *membros típicos* da categoria e a relevância das características do objeto para a categoria. O intuito dessa abordagem é construir os *protótipos* fazendo o "*download*" do conhecimento dos modelos CNN de classificação pré-treinados em uma estrutura semântica (*protótipo semântico*) que tenta representar o *significado semântico central* das categorias aprendidas pelo modelo CNN de classificação em questão.

Na Figura 4.1b é apresentado o fluxo dos principais processos que transformam a imagem de entrada, na assinatura do Descritor Semântico Global proposto. A abordagem de descrição semântica global de objetos apresentada, primeiramente *recupera* o protótipo da categoria baseado nas características observadas do objeto, e, em seguida utiliza o protótipo da categoria para encontrar as *diferenças semânticas* das características do objeto com relação ao *protótipo semântico* da categoria.

O processo de recuperação do protótipo da categoria (a partir das características extraídas do objeto) pode ser entendido como uma tarefa de *classificação semântica* que simula –e introduz nas CNN– o conceito de *categorização baseada em protótipos* (Rosch, 1978). O Descritor Semântico Global proposto constrói uma *representação semântica do objeto* (Ver Figura 4.1b 5)) que, devido à grande dimensionalidade, é reduzida para uma assinatura final que encapsula e preserva o significado presente na representação semântica inicial (Ver Figura 4.1b 6)).

O conjunto dos processos que devem ser desenvolvidos para atingir o objetivo geral da pesquisa podem ser delimitados e formalizados como:

1. **Modelagem e construção dos protótipos semânticos das categorias de objetos**. Modelar matematicamente a representação do protótipo semântico das categorias de objetos. Construir os protótipos semânticos das categoria de objetos a partir das características extraídas dos modelos CNN de classificação

pré-treinados (Na Figura 4.1a, construir *4)* a partir de *1)*). Avaliar a semântica encapsulada no protótipo da categoria usando o modelo matemático proposto.

2. **Recuperação do protótipo semântico da categoria**. Um método para recuperar o protótipo semântico da categoria do objeto baseado estritamente nas características extraídas do objeto (Ver passos *1), 2)* e *3)* da Figura 4.1b). Esse procedimento visa simular o comportamento humano nas tarefas de reconhecimento e classificação de objetos usando o *conceito de categorização baseado em protótipos*. O propósito desse processo é recuperar (dos protótipos calculados *a priori*) o protótipo mais provável que representa a categoria do objeto.

3. **Construção do Descritor Semântico Global**. Desenhar um modelo de descrição semântica baseada em protótipos para descrever globalmente os objetos. Modelar uma representação semântica do objeto (*assinatura*) que destaca as características distintivas do objeto dentro da categoria, enquanto introduz o significado contido no protótipo semântico da categoria (Ver passos *5)* e *6)* da Figura 4.1b).

A seguir se apresenta em detalhe cada um dos processos descritos anteriormente em forma de capítulos. Em cada capítulo são apresentadas as definições e os modelos propostos para cada processo, além dos resultados experimentais correspondentes.

Capítulo 5

Modelo Computacional do Protótipo

A finalidade da *semântica* consiste em decompor *o significado* em unidades menores, unitariamente chamadas *característica semântica*, permitindo segmentar o significado (Geeraerts, 1993). Como apresenta Geeraerts (1993). para o caso das palavras, essa decomposição permite diferenciar as palavras de significado parecido e as palavras de significado oposto. Essa ideia resultou ser extensiva para outras áreas de pesquisa na tentativa de representar o *significado semântico* de um determinado *conceito semântico* (Yan & Naphade, 2005).

Em Visão Computacional, a detecção de conceitos semânticos constitui, essencialmente, uma tarefa de classificação que determina se uma imagem possui relevância para um determinado *conceito semântico*. Os conceitos semânticos abrangem uma ampla gama de tópicos como os relacionados com objetos, cenas, eventos, etc (Yan & Naphade, 2005). Detectar, reconhecer, representar e descrever esses conceitos semânticos automaticamente em uma imagem, ainda constituem tarefas desafiadoras. Neste capítulo apresenta-se uma abordagem que visa *representar o significado* das características extraídas do conceito semântico *objeto*.

Motivados pela hipótese de que o ser humano constrói representações abstratas das categorias de objetos (protótipos), este trabalho baseou-se nos estudos da *Semântica Cognitiva* relacionados com a Teoria dos Protótipos como abordagem para construir essas representações semânticas das categorias de objetos. Ao observar membros de uma categoria de objeto (Figura 1.1a 1)), o ser humano constrói uma representação abstrata (Figura 1.1a 3)) que encapsula *o significado semântico central* da categoria de objeto aprendida (Murphy, 2004; McRae & Jones, 2013). Neste capítulo, apresenta-se uma abordagem que visa simular esse processo de aprendizagem de protótipos semân-

ticos. Especificamente, propõe-se construir –a partir de bancos de imagens anotadas de categorias de objetos– a representação semântica abstrata de uma categoria específica. Para atingir esse objetivo, a abordagem proposta baseia-se na hipótese de descomposição dos significados de Geeraerts (1993), mas em sentido inverso. Ou seja, a semântica do objeto pode ser representada a partir das características semânticas unitárias que o compõem, e do mesmo modo, a representação semântica abstrata da categoria dependerá dos significados semânticos dos objetos que a constituem.

Aliás, a modelagem e a construção da representação abstrata das categorias de objetos (os protótipos) foi baseada na visão prototípica desenvolvida por Rosch (Rosch & Mervis, 1975; Mervis & Rosch, 1981; Rosch, 1988; Geeraerts, 2010) para categorias que denominam objetos naturais. Com essa perspectiva, os membros estruturalmente mais salientes ou típicos (casos centrais claros) de uma categoria constituem o *protótipo* da categoria (Rosch, 1988; Geeraerts, 2010). A modelagem desse comportamento é complexa no domínio semântico, pois as características extraídas devem ser consideradas como *condições típicas* (Geeraerts, 2010) e não –somente– como características unitárias. Consequentemente, serão usados somente os *membros típicos* das categorias de objetos para a modelagem e construção do protótipo semântico proposto.

5.1 Representação Semântica

Nesta pesquisa pretende-se construir uma representação semântica que permita simular a estrutura semântica interna das categorias dos objetos. A *estrutura semântica* de uma categoria, ou seja, o *significado central/periférico*, está indubitavelmente relacionada com as diferenças de tipicidade e a presença de diferenças de saliência dos membros da categoria (*extensional non-equality*). Dentro dessa estrutura semântica o *protótipo abstrato* constitui a *média* da abstração de todos os objetos da categoria (Sternberg & Sternberg, 2016). O protótipo resume os membros mais representativos da categoria para determinadas características observadas. Consequentemente, os significados periféricos podem ser interpretados como aquelas características (ou conjunto de características) representadas pelos elementos que estão na periferia da categoria (membros periféricos), mas dentro das características representativas que definem à categoria.

A relevância relativa das características unitárias de uma categoria específica constitui outro aspecto importante da Teoria dos Protótipos. Essa relevância relativa permite que a representatividade de uma característica unitária possa variar quando se muda de categoria. A relevância unitária também permite que dentro de uma mesma categoria existam características mais relevantes que outras. A combinação das

características observadas do objeto e a sua correspondente relevância para a categoria permite o agrupamento de objetos em semelhança familiar (*intensional non-equality*). Essa abordagem justifica a posição semântica do objeto dentro da estrutura semântica da categoria e permite que objetos típicos sejam agrupados no centro semântico da categoria (próximos do *protótipo abstrato*) gerando, dessa maneira, uma organização prototípica dos membros da categoria.

Seja O um *universo de objetos*, $C = \{c_1, c_2, ..., c_n\}$ o conjunto finito de rótulos de categorias de objetos que particionam O, $O_{c_i} = \{o \in O : categoria(o) = c_i\}$ o conjunto de objetos que pertencem à i-ésima categoria $c_i \in C$ $(\forall i = 1, ..., n)$, e $F = \{f_1, f_2, ..., f_m\}$ o conjunto finito de características (*features*) distintivas dos objetos.

Definição 1. *Protótipo semântico.* O *significado semântico central* da i-ésima categoria $c_i \in C$, *protótipo semântico da categoria* c_i ou simplesmente *protótipo semântico*, o constitui a 3-*tupla* $P_i = (M_i, \Sigma_i, \Omega_i)$ onde $\forall i = 1, ..., n$; $\forall j = 1, ..., m$:

i) $M_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{im}]$ é um vetor m-dimensional não nulo onde μ_{ij} é o valor médio da j-ésima caraterística (f_j) extraída somente para *objetos típicos* da i-ésima categoria $c_i \in C$.

ii) $\Sigma_i = [\sigma_{i1}, \sigma_{i2}, ..., \sigma_{im}]$ é um vetor m-dimensional não nulo onde σ_{ij} é o desvio padrão da j-ésima característica extraída dos *objetos típicos* da categoria $c_i \in C$.

iii) $\Omega_i = [\omega_{i1}, \omega_{i2}, ..., \omega_{im}]$ é um vetor m-dimensional não nulo onde ω_{ij} é o valor de relevância da j-ésima característica (f_j) para a categoria $c_i \in C$.

Definição 2. *Protótipo abstrato.* O *centro semântico abstrato* da i-ésima categoria $c_i \in C$, elemento *mais prototípico* da categoria $c_i \in C$, *elemento ideal* da categoria $c_i \in C$ ou simplesmente *protótipo abstrato* da categoria c_i, é o vetor m-dimensional $M_i \in P_i = (M_i, \Sigma_i, \Omega_i)$ composto pelo valor esperado (valor médio) de cada uma das características dos objetos típicos da categoria $c_i \in C$, $\forall i = 1, ..., n$.

Definição 3. *Conjunto de protótipos semânticos.* O *conjunto de protótipos semânticos* do universo de objetos O, o banco de dados de protótipos semânticos que representa a O, o constitui o conjunto finito n-dimensional $P^O = \{P_1, P_2, ..., P_n\}$ onde P_i representa o *protótipo semântico* da i-ésima categoria $c_i \in C$, $\forall i = 1, ..., n$.

5.2 Distância Semântica

A abordagem de descrição semântica de objetos baseada em protótipos proposta (Vide os processos 4 - 5 na Figura 1.2) precisa de uma medida de similaridade para calcular a discrepância entre as características do objeto e as características típicas da categoria. Observa-se que a representação proposta para o protótipo semântico da categoria encapsula, no protótipo abstrato, o valor esperado das características típicas da categoria, já que foi construído exclusivamente com os objetos típicos da categoria. Consequentemente, a hipótese de descrição semântica proposta assume que a *distinção semântica* do objeto dentro da categoria dependerá de quão semelhante é o objeto com relação ao protótipo semântico da categoria. Mas, qual medida de similaridade usar para quantificar essa *distinção semântica* do objeto?

As métricas de distância L1 e L2 normalmente são usadas como métricas de similaridade entre vetores, mas sob a perspectiva semântica da Teoria dos Protótipos não constituem boas opções, pois assumem que todas as características unitárias do objeto possuem a mesma relevância. A Teoria do Protótipo expõe que: *i)* cada característica do objeto tem uma relevância relativa dentro da categoria; e *ii)* a relevância (ou saliência) de cada membro da categoria está de acordo com o quantidade e tipos de características presentes no objeto. Essa abordagem permite estabelecer um grau de prototipicidade de um elemento específico dentro da categoria (*extensional non-equality*). Nesta seção, generalizam-se algumas medidas de *distância entre estímulos* propostas na *Psicologia Experimental* (Vide Equações 3.1 e 3.4) com o objetivo de propor uma métrica de *distância semântica* (ou função de dissimilaridade) que quantifica a discrepância entre dois objetos (ou entre o objeto e o protótipo semântico) no contexto de uma categoria específica e com base nas características observadas do objeto.

Definição 4. *Distância semântica entre objetos.* Sejam os objetos $o_1, o_2 \in O$; F_{o_1}, F_{o_2} as características dos objetos o_1, o_2 respectivamente. Define-se como *distância semântica entre os objetos* o_1 e o_2 no domínio da categoria $c_i \in C$, à distância semântica $\delta(o_1, o_2)$ definida como:

$$\delta(o_1, o_2) = \sum_{j=1}^{m} |\omega_{ij}| \left| f_j^1 - f_j^2 \right|, \tag{5.1}$$

onde: $\omega_{ij} \in \Omega_i$, $f_j^1 \in F_{o_1}$, $f_j^2 \in F_{o_2}$, e $\Omega_i \in P_i; \forall j = 1...m; \forall i = 1...n$.

Observa-se que a *distância semântica entre objetos* proposta é uma generalização da *distância psicológica entre dois estímulos* do Modelo formal GCM (Ver Equação 3.1)

onde assume-se que: *i*) as características (estímulos) do objeto não são valores binários ($f_j \in \mathbb{R}$); *ii)* a relevância (custo de atenção) (ω_{ij}) de cada j-ésima característica unitária do objeto é forçada a ser estritamente positiva, mas não possui limite superior ($\sum \omega_{ij} \neq 1$). As restrições iniciais do Modelo GCM foram eliminadas com o objetivo de modelar a relevância da característica como os pesos aprendidos pelos modelos CNN de classificação.

Definição 5. *Distância prototípica.* Seja o objeto $o \in O$, F_o as características observadas do objeto o, e $P_i = (M_i, \Sigma_i, \Omega_i)$ o protótipo semântico da i-ésima categoria $c_i \in C$. Define-se como *distância prototípica* entre o objeto o e o protótipo semântico P_i no contexto da i-ésima categoria, à *distância semântica* $\delta(o, P_i)$ definida como:

$$\delta(o, P_i) = \sum_{j=1}^{m} |\omega_{ij}| \, |f_j - \mu_{ij}| \,, \tag{5.2}$$

onde: $\omega_{ij} \in \Omega_i$, $\mu_{ij} \in M_i$, $f_j \in F_o$; e $\Omega_i, M_i \in P_i$; $\forall j = 1, ..., m$; $\forall i = 1, ..., n$.

A *distância prototípica* proposta é uma generalização da distância semântica do Modelo formal MPM (Ver Equação 3.4) onde assume-se que as características do protótipo (μ_{ij}) não são as características de um elemento real da categoria, mas sim as características do elemento ideal construído (ou protótipo abstrato) da categoria ($M_i \in P_i$).

Definição 6. *Espaço métrico das características.* Seja F_{c_i} o conjunto não vazio de todas as características extraídas dos objetos da i-ésima categoria $c_i \in C$. A função de *distância semântica* $\delta : F_{c_i} \times F_{c_i} \to \mathbb{R}^+$ é uma *métrica* no domínio das características F_{c_i} pois satisfaz os axiomas: *não-negatividade*, *identidade de indiscernível*, *simetria* e *desigualdade triangular*. Consequentemente, (F_{c_i}, δ) constitui um *espaço métrico* ou *espaço métrico das características* da i-ésima categoria.

Demonstração. Sejam o_1, o_2, $o_3 \in O_{c_i}$ membros da i-ésima categoria $c_i \in C$; F_{o_1}, F_{o_2}, $F_{o_3} \in F_{c_i}$ as características observadas dos objetos o_1, o_2, o_3 respectivamente; e $f_j^k \in F_{o_k}$ a j-ésima característica unitária do elemento o_k; δ é uma *métrica* no domínio das características da i-ésima categoria, pois satisfaz os axiomas:

- $\delta(o_1, o_2) \geq 0$ *(não-negatividade)*
  Como todos os termos da Equação 5.1 são não negativos (≥ 0), $\delta(o_1, o_2) \geq 0$.

- $\delta(o_1, o_2) = 0 \Leftrightarrow o_1 = o_2$ *(identidade de indiscernível)*
  - $\delta(o_1, o_2) = 0 \to o_1 = o_2$. Se $\delta(o_1, o_2) = \sum_{j=1}^{m} |\omega_{ij}| \left| f_j^1 - f_j^2 \right| = 0$; $\forall \, |\omega_{ij}| \neq 0 \to \left| f_j^1 - f_j^2 \right| = 0 \to f_j^1 = f_j^2 \to o_1 = o_2$.

- $\delta(o_1, o_2) = 0 \leftarrow o_1 = o_2$. Se $o_1 = o_2 \rightarrow f_j^1 = f_j^2 \rightarrow \left|f_j^1 - f_j^2\right| = 0, \forall j = 1...m \rightarrow \delta(o_1, o_2) = 0$.

- $\delta(o_1, o_2) = \delta(o_2, o_1)$ *(simetria)*
  $\delta(o_1, o_2) = \sum_{j=1}^{m} |\omega_{ij}| \left|f_j^1 - f_j^2\right| = \sum_{j=1}^{m} |\omega_{ij}| \left|f_j^2 - f_j^1\right| = \delta(o_2, o_1)$

- $\delta(o_1, o_3) \leq \delta(o_1, o_2) + \delta(o_2, o_3)$ *(desigualdade triangular)*
  $\delta(o_1, o_2) + \delta(o_2, o_3) = \sum_{j=1}^{m} |\omega_{ij}| \left|f_j^1 - f_j^2\right| + \sum_{j=1}^{m} |\omega_{ij}| \left|f_j^2 - f_j^3\right| = \sum_{j=1}^{m} |\omega_{ij}| \left(\left|f_j^1 - f_j^2\right| + \left|f_j^2 - f_j^3\right|\right)$ por propriedade do valor absoluto $\left|f_j^1 - f_j^2\right| + \left|f_j^2 - f_j^3\right| \geq \left|f_j^1 - f_j^3\right| \rightarrow \delta(o_1, o_2) + \delta(o_2, o_3) \geq \delta(o_1, o_3)$.

□

Observe também que se $E \subseteq O_{c_i}$ é um subconjunto da i-ésima categoria, então $\delta(E) = \sum \delta(o, P_i)$, $\forall o \in E$. Consequentemente, a distância prototípica proposta satisfaz as propriedades: *i)* conjunto vazio nulo (*null empty set*): $\delta(\varnothing) = 0$; *ii)* aditividade contável (*countable additivity*): para todas as coleções contáveis $\{E_k\}_{k=1}^{\infty}$ de conjuntos disjuntos em pares que pertencem a E, $\delta\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \delta(E_k)$ (É simples provar esta propriedade usando o método de indução matemática).

Corolário 1. A função de distância prototípica do conjunto de características do objeto F_{c_i} até a linha do número real estendida, $\delta : F_{c_i} \rightarrow \mathbb{R}^+$, é uma medida. Consequentemente, (F_{c_i}, δ) é um espaço mensurável.

Como a definição do domínio das características (F_{c_i}, δ) é um espaço mensurável, podemos usar a generalização da desigualdade de Chebyshev (Chebyshev, 1867) para definir o limite da representação do protótipo semântico proposto. Chebyshev (1867) afirmou que uma variável aleatória escalar ξ com distribuição Pr pode diferir de sua média $\mu \in \mathbb{R}$ em mais de $\lambda \in \mathbb{R} > 0$ desvios padrão $\sigma \in \mathbb{R} > 0$, com uma probabilidade que sempre satisfaz a expressão: $\Pr(|\xi - \mu| \geq \lambda\sigma) \leq \min(1, \frac{1}{\lambda^2})$.

Saw et al. (1984) e Stellato et al. (2017) abordaram o problema de formular uma desigualdade empírica de Chebyshev com N i.i.d amostras de uma distribuição desconhecida Pr, sua média empírica μ_N e o desvio padrão empírico σ_N. Saw et al. (1984) e Stellato et al. (2017) derivaram uma desigualdade de Chebyshev relacionada com a amostra $(N+1)$-ésima. A desigualdade de Chebyshev Multivariada (Multivariate Chebyshev inequality)(Stellato et al., 2017) pode definir o limite de um conjunto elipsoidal centrado na média da amostra.

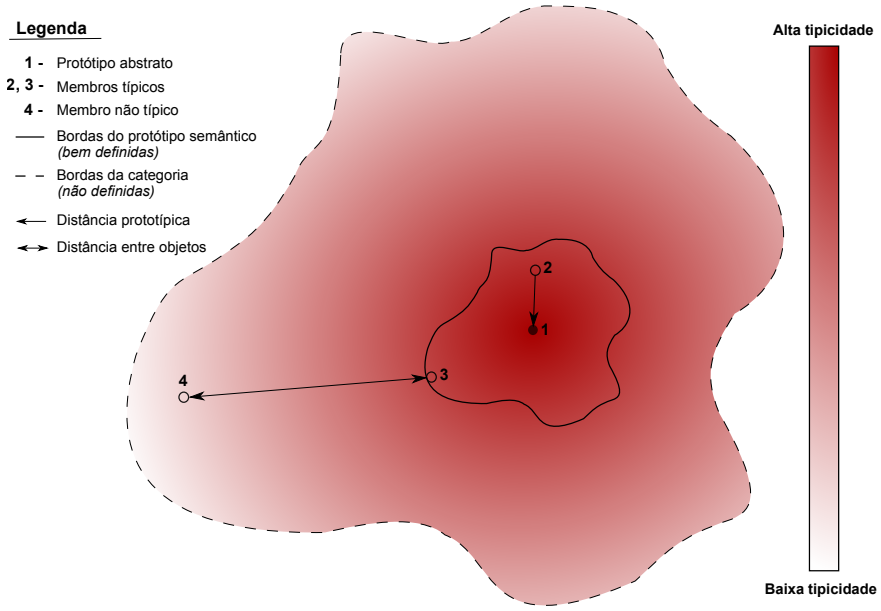


Figura 5.1: *Resumo visual do modelo CPM proposto.* O modelo CPM proposto está constituído pela representação do protótipo semântico $P_i = (M_i, \Sigma_i, \Omega_i)$ e pelas relações de distância semântica (δ) existentes entre os membros da categoria. Mostra-se visualmente a representação semântica esperada da estrutura interna da categoria, e o conjunto de definições construídas para a formalização do modelo CPM proposto. Fonte: Elaborado pelo autor.

Definição 7. *Bordas do protótipo semântico.* Seja (F_{c_i}, δ) o espaço métrico das características dos objetos da i-ésima categoria $c_i \in C$. Seja $E \subseteq F_{c_i}$ o subconjunto das características extraídas de apenas os *objetos típicos* da i-ésima categoria, $N = |E|$ a quantidade de elementos que compõem E, e F_o as características extraídas do objeto $o \in O_{c_i}$. Definem-se *fracamente* como as bordas do protótipo semântico $P_i = (M_i, \Sigma_i, \Omega_i)$, o vetor limiar $\vec{\lambda_i} = [\lambda_{i1}, \lambda_{i2}, ..., \lambda_{im}]$ que cumpre com a expressão:

$$\Pr(|f_j - \mu_{ij}| \geq \lambda_{ij}\sigma_{ij}) \leq \min(1, \frac{1}{\lambda_{ij}^2}), \tag{5.3}$$

onde $f_j \in F_o$, $\mu_{ij} \in M_i$ e $\sigma_{ij} \in \Sigma_i, \forall i = 1...n; \forall j = 1...m.$. Finalmente, dado um limiar de probabilidade, é possível calcular um vetor limiar $\vec{\lambda_i}$ e construir um conjunto elipsoidal de confiança a partir da média da amostra e da covariância do conjunto de objetos típicos (Uma definição mais forte de *bordas do protótipo semântico* pode ser construída usando - completamente - as definições de Stellato et al. (2017)).

A Figura 5.1 apresenta um resumo visual do conjunto de definições que constituem o Modelo Computacional do Protótipo (*Computational Prototype Model (CPM)*)

proposto. Apresenta-se a representação esperada da estrutura semântica interna da categoria baseado no modelo CPM proposto [*Protótipo Semântico* (Definição 1) + *Distância Semântica* (Definições 4, 5)]. O modelo CPM proposto, procura respeitar alguns conceitos importantes da Teoria dos Protótipos: *i)* os limites do protótipo semântico da categoria se encontram bem definidos com o vetor $\Sigma_i \in P_i$; *ii)* as bordas da categoria são difusas (não estão definidas) porque o protótipo semântico proposto não é construído com todos os elementos de categoria (somente com membros típicos); *iii)* o protótipo abstrato da categoria constitui o elemento que representa o centro semântico da categoria; *iv)* a representatividade dos objetos (tipicidade) dentro da categoria tenta ser simulada com a *distância semântica* proposta.

5.3 Construção do Protótipo Semântico

A natureza da representação do protótipo semântico proposto (Ver Definição 1) permite que o protótipo seja facilmente calculado usando qualquer modelo computacional com a capacidade de: *i)* extrair as características do objeto em imagens (Fo); e *ii)* aprender o valor de relevância unitária (ω_{ij}) de cada j-ésima característica do objeto relativo à i-ésima categoria. Outro aspecto necessário para calcular o protótipo semântico proposto é conhecer o valor de tipicidade dos elementos dentro da categoria. Consequentemente, para construir corretamente os protótipos semânticos com a abordagem proposta, é preciso ter um banco de dados de imagens de objetos com as anotações da pontuação de tipicidade de cada objeto. Por outro lado, pelos pré-requisitos específicos desta pesquisa, o modelo de extração de características usado para a construção dos protótipos semânticos deve ter, além dos requerimentos anteriores, também a habilidade de classificar o objeto.

5.3.1 Seleção do modelo de classificação

O paradigma de aprendizagem profunda permitiu construir modelos CNNs de classificação altamente robustos para lidarem com as mudanças semânticas das categorias. Essa característica permitiu que esses modelos ultrapassarem –pela primeira vez– o desempenho humano em tarefas de classificação de imagens a grande escala; fato que justifica o sucesso dos modelos CNN em tarefas de classificação. O excelente desempenho dos modelos CNNs de classificação fundamenta-se na construção de características de maior abstração, complexidade e de maior poder discriminativo. Essas propriedades propiciaram o uso dos modelos CNN de classificação na extração de características

discriminativas da imagem, fato que demonstrou ter um bom desempenho em tarefas de compreensão semântica da imagem.

A análise desenvolvida por Sun et al. (2017) (Figura A.4), no processo evolutivo dos modelos CNNs de classificação no desafio de ImageNet, confirmou que os novos modelos CNN de classificação sempre superaram em desempenho aos modelos anteriores. Sun et al. (2017) destacaram que, contrário à acelerada evolução dos modelos CNN de classificação, os dados de treinamento desses modelos CNN praticamente permaneceram constantes no mesmo período de tempo.

Esses pressupostos motivaram o processamento dos protótipos semânticos das categorias de objetos usando os modelos CNN de classificação para a extração das características das imagens nesses bancos de dados. Várias são as vantagens dessa abordagem: *i)* realizar a tarefa de extração das características do objeto usando modelos CNN de classificação já pré-treinados; *ii)* conseguir portabilidade e escalabilidade na abordagem de descrição de objetos proposta, pois –ao basear-se em modelos CNNs de classificação pré-treinados– qualquer evolução ou contribuição nessa área de classificação também contribuirá na evolução da metodologia de descrição semântica proposta; *iii)* usar bancos de dados de imagens que não mudam com frequência permite construir protótipos semânticos invariantes para cada categoria de objeto. Ou seja, o significado semântico encapsulado nos protótipos semânticos calculados com um modelo CNN para um determinado banco de dados de imagens, permanecerá constante sempre que o banco de dados em questão não seja acrescentado com novas imagens ou categorias de objetos.

5.3.2 Características e Relevância das Características

Os modelos CNN de classificação de imagens utilizam –comumente– a *função softmax* como *função de ativação* da camada superior da rede convolucional. O uso da função softmax nessa camada totalmente conetada (*full conected layer*) permite aprender a relevância de cada j-ésima característica unitária para cada i-ésima categoria de objeto aprendida. Lake et al. (2015) mostraram que a saída da última camada dos modelos CNN de classificação pode ser usada como um sinal de *quão típica* é uma imagem de entrada. Os autores mostraram que o valor da tipicidade do objeto dentro da categoria pode ser proporcional à magnitude da resposta da classificação à categoria de interesse. Outros estudos (Khaligh-Razavi & Kriegeskorte, 2014; Cichy et al., 2017) concluíram que a combinação ponderada das características da última camada completamente conectada dos modelos CNN pode explicar completamente o córtex temporal inferior do cérebro humano.

Baseado nesses pressupostos, e no pré-requisito do presente trabalho de usar as mesmas características para classificar e para descrever objetos, a construção do protótipo semântico proposto usa as características extraídas das imagens com os modelos CNNs de classificação. Formalmente, dado um modelo CNN de classificação pré-treinado, assume-se como as *características do objeto* aquelas extraídas da camada superior totalmente conectada do modelo CNN em questão. Definem-se também como *valores de relevância* das características extraídas do objeto para a i-ésima categoria $c_i \in C$, aos pesos aprendidos (Ω_i, b_i) pela função *softmax* no *i-ésimo* neurônio de saída da camada superior totalmente conectada do modelo CNN; $\forall i = 1, ..., n$.

Definição 8. *Protótipo semântico convolucional.* O *protótipo semântico convolucional* da categoria $c_i \in C$ representa-se pela 4-tupla $P_i = (M_i, \Sigma_i, \Omega_i, b_i)$ onde M_i, Σ_i são calculados usando as características dos objetos da i-ésima categoria c_i, extraídas da camada completamente conectada (*fully convolutional layer*) de uma rede neuronal convolucional. Os elementos Ω_i, b_i da 4-tupla P_i representam os i-ésimos *valores de relevância* aprendidos pela camada *softmax* para a i-ésima categoria $c_i, \forall i = 1, ..., n$.

Doravante, é referido como *protótipo semântico* da i-ésima categoria $c_i \in C$ ao *protótipo semântico convolucional* representado pela 4-tupla $P_i = (M_i, \Sigma_i, \Omega_i, b_i)$, onde os elementos membros da tupla são calculados usando as características extraídas e os pesos aprendidos pelos modelos CNN de classificação pré-treinados. Observe-se que com essa abordagem são garantidas *representações semânticas únicas* para os protótipos semânticos de cada categoria de objeto calculados com os modelos CNN de classificação. Essa propriedade de *unicidade* dos protótipos semânticos projetados é garantida pelos componentes Ω_i, b_i da 4-tupla P_i, pois a função *softmax* na camada totalmente conectada do modelo CNN, garante que esses pesos aprendidos sejam únicos para cada i-ésima categoria do banco de dados.

5.3.3 Algoritmo de Construção

Nesta pesquisa, os protótipos semânticos propostos foram projetados usando como ponto de partida o conhecimento aprendido pelos modelos CNN de classificação pré-treinados. A abordagem usada justifica-se pelo fato de que os modelos CNN de classificação, de forma análoga à memória humana (Fuster, 1997), fazem associações que mantêm o conhecimento aprendido nas suas estruturas de conexão entre os neurônios. A abordagem de representação semântica proposta faz o *download* do conhecimento desses modelos CNN de classificação pré-treinados em uma estrutura semântica (pro-

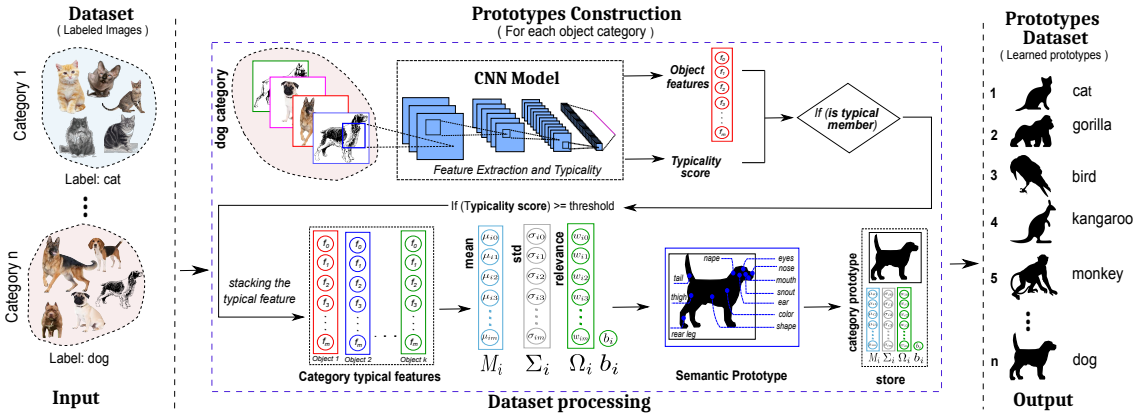


Figura 5.2: *Construção off-line do banco de dados de protótipos semânticos.* Dado um conjunto de dados de imagens rotuladas, para cada categoria de objetos presente no conjunto de dados, calcula-se a representação de protótipo semântico proposto usando o Algoritmo 1.

Algoritmo 1 Construção do Protótipo Semântico

Entrada: Modelo CNN Λ, banco de dados de objetos O, categoria c_i
Saída: Protótipo Semântico da Categoria (P_i)
$O_{c_i} \leftarrow \{o \in O : categoria(o) = c_i\}$
$features\_block \leftarrow \{\}$
for $o \in O_{c_i}$ **do**
    if o $is\_typical$ **then**
        $F_o \leftarrow \Lambda.features\_of(o)$
        $features\_block \leftarrow features\_block \cup F_o$
$\Omega_i, b_i \leftarrow \Lambda.sofmax\_weight\_learned\_of(c_i)$
$M_i, \Sigma_i \leftarrow compute\_stats(features\_block)$
return $(M_i, \Sigma_i, \Omega_i, b_i)$

tótipo semântico) que visa representar o significado semântico central das categorias de objetos aprendidas.

O Algoritmo 1 detalha como calcular o protótipo semântico proposto para uma categoria específica. Dado um banco de dados rotulado de imagens de objetos (O), e um modelo CNN de classificação pré-treinado (Λ); para cada i-ésima categoria de objeto no conjunto de dados (O) usa-se o Algoritmo 1 para calcular o protótipo semântico correspondente à i-ésima categoria de objeto $(P_i = (M_i, \Sigma_i, \Omega_i, b_i))$.

A Figura 5.2 mostra os passos e os conceitos principais do Algoritmo de construção de protótipos semânticos proposto. A Figura 5.2 resume visualmente o processamento *off-line* dos protótipos semânticos correspondentes às categorias de um determinado banco de dados de imagens (O) seguindo o fluxo de passos apresentado

no Algoritmo 1. Para cada i-ésima categoria ($\forall c_i \in C, i = 1...n$) do banco de dados de imagens (O), agrupam-se as imagens rotuladas que pertencem àquela i-ésima categoria de objeto (O_{c_i}). Em seguida são extraídas, usando um modelo CNN de classificação, as características correspondentes às imagens daquela categoria de objeto. Finalmente, se calcula cada elemento que compõe a 4-tupla do protótipo semântico $P_i = (M_i, \Sigma_i, \Omega_i, b_i)$; baseado nas características extraídas dos *objetos típicos* da categoria e nos pesos aprendidos (Ω_i, b_i) para aquela categoria de objeto (c_i). Como resultado final obtêm-se o banco de dados de protótipos semânticos (ou *conjunto de protótipos semânticos* (Ver Definição 3)) correspondente àquele banco de dados de imagens (O) para o modelo CNN de classificação selecionado (Λ). Observa-se que o *conjunto de protótipos semânticos* resultante desse processamento *off-line* é usado como *conhecimento prévio* no modelo de descrição semântica baseado em protótipos proposto (Vide Figura 1.2).

É válido esclarecer que a construção do *protótipo semântico* proposto é realizada usando como modelo de extração de características uma instância de um modelo CNN de classificação pré-treinado (Ver Figura 5.2). Mas, o Algoritmo 1 pode ser usado com qualquer modelo de classificação que possua os pré-requisitos apresentados na Seção 5.3.

5.3.4 Representação gráfica do protótipo

A abordagem prototípica visa construir uma representação abstrata do conhecimento das categorias de objetos. Essa representação captura as propriedades estruturalmente mais salientes da categoria e define uma estrutura bem delimitada que constitui o protótipo semântico da categoria. Representar visualmente o protótipo permite apresentar um resumo visual das definições semânticas representativas (ou características típicas) da categoria. Por exemplo, o esboço apresentado na Figura 1.1a (3) representa visualmente o protótipo que encapsula o conjunto de características típicas (ou representativas) dos membros da categoria *cão* representados na Figura 1.1a (1).

Conseguir uma representação visual *no domínio da imagem* do protótipo semântico modelado (P_i) não constitui uma tarefa ligeira. Uma abordagem de visualização seria construir um modelo DNN (Deconvolutional Neural Network) para realizar a *engenharia reversa* dos parâmetros de entrada do modelo CNN de classificação usado para a construção do protótipo. Essa abordagem normalmente constrói a imagem resultante a partir das características e aprende pesos próprios; mas, pelas particularidades do protótipo semântico proposto, é um resultado não desejado neste contexto. Algumas abordagens (Wohlhart et al., 2013; Li et al., 2018) aprenderam representações

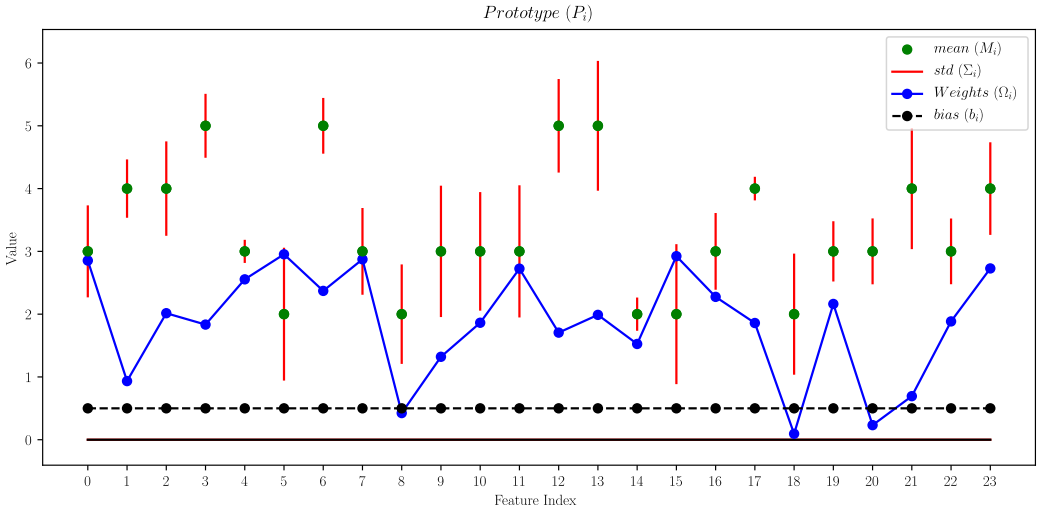


Figura 5.3: *Representação gráfica do protótipo semântico P_i modelado para a categoria c_i.* Os vetores m-dimensionais $M_i, \Sigma_i, \Omega_i, b_i$ que constituem os elementos da tupla P_i representam-se nas cores verde, vermelho, azul e preto respectivamente (Ver em cores). Fonte: Elaborado pelo autor.

de protótipos diretamente no domínio da imagem; e, consequentemente, a visualização do protótipo é simplesmente a imagem aprendida. Essas abordagens exigem gastos computacionais consideráveis na aprendizagem da visualização dos protótipos calculados. Wohlhart et al. (2013) introduziu a aprendizagem de representações de imagens de protótipos no processo de *backpropagation* da rede construída pelo autor. Por outro lado, Li et al. (2018) usou uma arquitetura profunda de codificador-decodificador para aprender a visualização de protótipos. Como a representação de protótipo semântico proposto é construída a partir de modelos de classificação da CNN pré-treinados, foi usada outra abordagem para visualizar a representação dos protótipos semânticos calculados.

Binder et al. (2016) propôs uma visualização circular dos vetores compostos pela média dos atributos semânticos correspondentes a categorias de substantivos de objetos. Consequentemente, uma abordagem simples para visualizar a representação do protótipo semântico proposto é visualizar os vetores m-dimensionais que compõem a tupla do protótipo semântico.

A Figura 5.3 mostra a representação gráfica do protótipo semântico $P_i = (M_i, \Sigma_i, \Omega_i, b_i)$ da i-ésima categoria $c_i \in C$. Mostram-se os vetores m-dimensionais que constituem os elementos do protótipo semântico proposto. O valor aprendido b_i apresenta-se como o vetor m-dimensional constituído pelos *bias* uniformes de valor $\frac{b_i}{m}$ (preto). O vetor m-dimensional M_i (verde) representa o valor médio das características extraídas dos membros típicos da i-ésima categoria. O vetor m-dimensional

Ω_i (azul) constitui os valores de relevância de cada característica unitária dentro da i-ésima categoria. O vetor m-dimensional Σ_i (linhas vermelhas) constitui o desvio padrão do valor esperado para as características típicas da categoria. Observa-se que a exclusividade do protótipo semântico proposto é garantida pelo vetor de relevância Ω_i, aprendido especificamente para a i-ésima categoria no processo de treinamento do modelo CNN de classificação.

5.4 Significado Semântico do Objeto

Várias pesquisas de neurociência cognitiva estudaram o efeito do *significado semântico* na tarefa de reconhecimento de objetos (Tulving, 2007; Martin, 2007; Collins & Curby, 2013). Alguns resultados mostraram que quando um objeto foi previamente associado a algum tipo de significado semântico no cérebro, as pessoas são mais propensas a identificar corretamente o objeto (Tulving, 2007; Martin, 2007). Collins & Curby (2013) mostraram que as associações semânticas permitem um reconhecimento muito mais rápido do objeto, mesmo quando a tarefa de reconhecimento de objetos se torna cada vez mais difícil (mudanças nos pontos de vista, oclusão de partes do objeto, etc.). Assim, esses trabalhos concluíram que as associações semânticas do cérebro baseadas no significado semântico do objeto permitem o reconhecimento mais rápido de objetos (Tulving, 2007; Martin, 2007; Collins & Curby, 2013) .

Além disso, o fato de que alguns modelos CNN de classificação (por exemplo, ResNet (He et al., 2016)) superaram o desempenho relatado por humanos (5.1% (Russakovsky et al., 2015a)) em tarefas de classificação de objetos visuais em grande escala, geraram alguns estudos cognitivos (Yamins et al., 2014; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cichy et al., 2017) para pesquisar as possíveis ligações existentes entre o funcionamento dos modelos CNN e o sistema visual no cérebro humano. Cichy et al. (2017) sugeriram que as CNNs realizam representações de arranjos espaciais como aquelas realizadas pelos humanos. Khaligh-Razavi & Kriegeskorte (2014) concluíram que a combinação ponderada de características na última camada completamente conectada dos modelos CNN pode explicar completamente o córtex temporal inferior no cérebro humano. A presente pesquisa, baseia-se nesses fundamentos teóricos para modelar uma representação do significado semântico dos objetos usando os modelos CNN.

O significado semântico das características observadas de um objeto depende fortemente da relevância (Ω) que cada característica unitária possui para cada categoria (por exemplo, na categorização por *cores* a característica *altura* não possui re-

levância). O significado semântico do objeto varia, assim como a sua saliência (ou valor de tipicidade), desde o contexto de cada i-ésima categoria. Esse comportamento é consequência do fato de que cada i-ésima categoria aprendida atribui diferentes valores de relevância (ω_{ij}) às j-ésimas características unitárias de seus membros, $\forall i = 1, ..., n; \forall j = 1, ..., m$.

Definição 9. *Valor semântico do objeto.* Seja $F_o = \{f_1, f_2, ..., f_m\}$ o conjunto de características observadas de um objeto $o \in O$. O *significado semântico* das características observadas do objeto F_o para a categoria $c_i \in C$, *valor resumo* das características observadas F_o, ou simplesmente *valor semântico* do objeto o para a categoria c_i; define-se pelo valor abstrato:

$$z_o = \sum_{j=1}^{m} \omega_{ij} f_j + b_i \tag{5.4}$$

onde: $\omega_{ij} \in \Omega_i,\ f_j \in F_o$, e $\Omega_i, b_i \in P_i;\ \forall j = 1, ..., m;\ \forall i = 1, ..., n$.

Note-se que a proposta de valor semântico do objeto (ou significado semântico do objeto) é exatamente o mesmo valor usado para categorizar objetos na camada *softmax* dos modelos CNN de classificação.

Definição 10. *Valor semântico da categoria.* O *significado semântico* esperado dos objetos da i-ésima categoria $c_i \in C$, *valor resumo* do protótipo semântico que representa à i-ésima categoria $P_i = (M_i, \Sigma_i, \Omega_i, b_i)$, ou simplesmente *valor semântico da categoria* c_i, define-se pelo valor abstrato:

$$\hat{z}_i = \sum_{j=1}^{m} \omega_{ij} \mu_{ij} + b_i \tag{5.5}$$

onde: $\omega_{ij} \in \Omega_i,\ \mu_{ij} \in M_i$, e $\Omega_i, M_i, b_i \in P_i;\ \forall j = 1, ..., m;\ \forall i = 1, ..., n$.

5.4.1 O Significado semântico e a distância semântica

O significado semântico (valor semântico) de um objeto para uma categoria específica pode ser facilmente calculado – no espaço semântico das características extraídas usando um modelo CNN de classificação – usando apenas as ativações da última camada completamente conetada do modelo CNN em questão (camada *softmax*)(Ver Equação 5.4). Por outro lado, o *protótipo semântico convolucional* da i-ésima categoria pode ser calculado de maneira simples usando o Algoritmo 1 e as características

extraídas com o modelo CNN selecionado. Mas, no espaço das características convolucionais, pode se estabelecer alguma relação entre o *significado semântico do objeto* e as relações de distância semântica definidas entre *objeto-objeto* (Ver Definição 4) e *objeto-protótipo semântico* (Ver Definição 5)? Sem perda de generalidade, pode se entender que a *distância semântica* entre *objeto-objeto* e *objeto-protótipo semântico* no contexto da i-ésima categoria, é a diferença entre os *significados semânticos* correspondentes. Por exemplo, o protótipo semântico da i-ésima categoria (P_i) e as características extraídas dos objetos (F_o) da i-ésima categoria pertencem ao mesmo domínio m-dimensional de características. Consequentemente, a *diferença semântica* entre o protótipo semântico e o objeto (Definição 5) pode ser entendida como a soma das diferenças absolutas entre os valores das características unitárias que compõem seus valores semânticos. Ou seja, usando as Equações 5.4 e 5.5 obtém-se a expressão: $|z_o - \hat{z}_i| \approx \sum_{j=1}^{m} |(\omega_{ij} f_j + b_i) - (\omega_{ij}\mu_{ij} + b_i)| = \sum_{j=1}^{m} |\omega_{ij} f_j - \omega_{ij}\mu_{ij}| = \sum_{j=1}^{m} \omega_{ij} |f_j - \mu_{ij}|$, é exatamente igual à *distância prototípica* (Ver Definição 5) quando $\omega_{ij} \geq 0$, $\forall \omega_{ij} \in \Omega_i$, $\forall i = 1, ..., n; \forall j = 1, ..., m$. Analogamente, pode se estabelecer uma relação semelhante entre a *distância semântica objeto-objeto* e os valores semânticos dos objetos em análise.

5.5 Organização prototípica da categoria

A construção do *protótipo semântico* da categoria converte às categorias de objetos em *categorias prototípicas*. Consequentemente, a abordagem prototípica permite representar o conhecimento da estrutura interna da categoria mediante a *organização prototípica* dos elementos da categoria. A modelagem da *organização prototípica* do conhecimento de cada i-ésima categoria foi baseada nos quatro *efeitos prototípicos* (Geeraerts, 1997, 2010): *extensional non-equality*, *intensional non-equality*, *extensional non-discreteness*, *intensional non-discreteness*.

A modelagem do significado central/periférico (conceito *non-equality*) de uma categoria de objeto tem importância crítica porque define a relevância (ou saliência) de um objeto específico dentro da categoria. Esse problema é consequência da existência de exemplares ou membros da categoria que demonstram ter diferenças de tipicidade e de saliência na categoria (nível *extensional*). Ou seja, dependendo das características semânticas observadas, existem membros mais representativos (ou típicos) que outros dentro da categoria. Essas diferenças de tipicidade permitem que o cérebro humano –de forma inata– agrupe os elementos em *semelhança familiar* dentro da categoria (Ver exemplo na Figura 3.1) e constitui um comportamento justificado pelo efeito prototípico *non-equality* no nível de definição (*intensional*).

Visualizar a posição semântica de cada elemento de uma categoria com respeito ao centro semântico da categoria (ou protótipo abstrato), constitui uma abordagem simples para observar a estrutura semântica interna dessa categoria. A dificuldade que possui essa abordagem é que a visualização da estrutura interna da categoria é inviável no espaço m-dimensional das características (sob a perspectiva dos fundamentos teóricos usados nesta pesquisa) sem o uso de técnicas de descarte de características para reduzir a dimensionalidade. Por essa razão são usadas técnicas de topologia como abordagem alternativa para visualizar a estrutura interna da categoria, e mostrar que o modelo CPM proposto pode simular a organização prototípica dos elementos dentro da estrutura semântica interna da categoria.

Sejam (F_{c_i}, δ) e $(\mathbb{R}^2, L_1)$ *espaços métricos*; e ρ a função que mapeia as características do objeto para o espaço métrico $(\mathbb{R}^2, L_1)$ usando o *valor semântico* e a *distância prototípica* do objeto. Ou seja, ρ é a *aplicação não expansiva* entre esses espaços métricos $\rho : F_{c_i} \to \mathbb{R}^2 \mid \rho(o \in O_{c_i}) = \rho(F_o) = p(z_o, \delta(o, P_i))$; onde F_o são as características do objeto, z_o representa o *valor semântico* do objeto; $\delta(o, P_i)$ representa a *distância prototípica*; $p(x, y)$ representa um ponto no domínio $\mathbb{R}^2$ nas coordenadas (x, y); e l_1 representa a Soma da Diferença Absoluta (*Sum of Absolute Difference (SAD)*) no domínio $\mathbb{R}^2$.

Proposição 1. *Relação entre as métricas δ e L_1.* Sejam os objetos $o_1, o_2 \in O_{c_i}$. Se $p_1 = \rho(o_1)$, $p_2 = \rho(o_2)$ são os correspondentes pontos mapeados no espaço métrico $(\mathbb{R}^2, L_1)$, então a relação entre as métricas dos espaços métricos (F_{c_i}, δ) e $(\mathbb{R}^2, l_1)$ pode ser expressada como:

$$\delta(o_1, o_2) \leq l_1(p_1, p_2) \leq 2\delta(o_1, o_2). \tag{5.6}$$

Demonstração. $L_1(p_1, p_2) = L_1(\rho(o_1), \rho(o_2)) = |z_1 - z_2| + |\delta_1 - \delta_2|$. Usando as Definições 4, 5, 9, 10 a expressão anterior pode ser expressada como: $L_1(p_1, p_2) = \left|\sum_{j=1}^{m} \omega_{ij}(f_j^1 - f_j^2)\right| + \left|\sum_{j=1}^{m} |\omega_{ij}| \left(\left|f_j^1 - \mu_{ij}\right| - \left|f_j^2 - \mu_{ij}\right|\right)\right|$. Finalmente, usando a propriedade da função valor absoluto $|a - b| - |c - b| \leqslant |a - c|$, é simples demonstrar a desigualdade $\delta(o_1, o_2) \leq l_1(p_1, p_2) \leq 2\delta(o_1, o_2)$, $\forall j = 1...m$; $\forall i = 1...n$. □

Consequentemente, para todo $F_{o_1}, F_{o_2} \in F_{c_i}$ e $\forall \varepsilon > 0$, sempre existe $\exists \varphi = \frac{\varepsilon+1}{2} > 0$ de modo que se $\delta(o_1, o_2) < \varphi \Rightarrow L_1(\rho(o_1), \rho(o_2)) < \varepsilon$. Ou seja, a função $\rho : F_{c_i} \to \mathbb{R}^2$ *é uma função contínua.* As funções contínuas entre espaços métricos preservam as propriedades dos espaços métricos e permitem descrever a deformação de um espaço métrico em outro. O anterior significa que os pontos na vizinhança de um objeto mapeado no

domínio $\mathbb{R}^2$, também pertencem à vizinhança do objeto no domínio F_{c_i} quando são mapeados novamente com a função inversa. Ou seja, $\rho(o_1) = p_1 \rightarrow \forall p \in \{vizinhana\ de\ p_1\}$, $\rho^{-}(p) \in \{vizinhana\ de\ o_1\}$. Assim, o comportamento observado (em termos da métrica de distância L_1) no espaço métrico $(\mathbb{R}^2, L_1)$ preserva-se (em termos da métrica de distância δ) no espaço métrico m-dimensional das características (F_{c_i}, δ).

Finalmente, visualizar a estrutura interna da i-ésima categoria resume-se à visualização no espaço métrico $(\mathbb{R}^2, L_1)$ de cada elemento da categoria $c_i \in C$ mapeado através da função ρ. Se como resultado da projeção da i-ésima categoria $(\rho(c_i))$, a categoria mapeada for organizada em torno do protótipo mapeado $\rho(P_i)$ baseado na tipicidade visual de seus membros; então o modelo proposto simula a organização prototípica dos elementos dentro da categoria c_i. Observa-se que essa projeção procura preservar a posição do objeto (distância com relação ao protótipo) em ambos os espaços, baseado nas métricas de distância correspondentes. Preservar a posição semântica no espaço métrico $(\mathbb{R}^2, L_1)$ significa que os objetos visualmente típicos continuam sendo encontrados próximos do protótipo $\rho(P_i)$ e os menos representativos permanecem distantes do centro semântico abstrato (protótipo abstrato) da categoria mapeada.

5.6 Experimentos e Resultados

Nesta seção são apresentados os resultados experimentais obtidos na avaliação do modelo CPM proposto. Com os experimentos realizados analisou-se o poder interpretativo do modelo CPM proposto (*protótipo semântico + distância semântica*) e a sua capacidade de capturar a relevância visual das imagens de objetos. Os experimentos realizados visam mostrar a habilidade da modelagem semântica proposta para: *i)* capturar, com o protótipo semântico, o significado semântico central de uma categoria de objeto específica; *ii)* simular, de forma comparável ao ser humano, que elementos visualmente típicos das categorias de objetos sejam organizados próximos (com base na métrica de distância semântica proposta) ao protótipo abstrato da categoria (organização prototípica da estrutura semântica interna da categoria).

5.6.1 Bancos de Dados e Modelos Selecionados.

Os experimentos apresentados foram realizados nos bancos de dados que seguem:

- *MNIST dataset (Lecun et al., 1998).* As poucas categorias (10 categorias) presentes nesse banco de dados e a particularidade de que todas as imagens de dígitos

estão cortadas, centradas e normalizadas, propiciaram sua escolha como o banco de dados piloto nos experimentos realizados.

- *CIFAR-10 (Krizhevsky & Hinton, 2009).* Um banco de imagens com características similares ao banco de dados MNIST, mas composto por imagens de objetos. O conjunto de dados consiste em imagens de objetos coloridas de 32×32 dimensões. O banco de imagens está composto por 10 categorias divididas em 50.000 imagens de treinamento e 10.000 imagens de teste.

- *CIFAR-100 (Krizhevsky & Hinton, 2009).* O banco de dados está conformado por imagens do mesmo tamanho e formato que o banco de dados CIFAR-10, mas contém 100 categorias de objetos.

- *ILSVRC 2014 (Russakovsky et al., 2015a).* O banco de dados ImageNet foi selecionado como banco de dados de objetos reais. O banco de imagens possui 1000 categorias de objetos reais e mais de um milhão de imagens com anotações que delimitam somente a região do objeto. Usaram-se essas regiões delimitadas para reduzir, na representação do protótipo semântico (P_i) da i-ésima categoria, o ruído adicionado dos elementos que não pertencem à i-ésima categoria e que aparecem em cada imagem.

Para cada banco de dados de imagens usado foi selecionado um modelo CNN de classificação para a extração de características e a construção do banco de protótipos semânticos correspondente. Com essa finalidade, foram construídos e treinados dois modelos CNN pouco profundos – modelos *simples-MNIST* (Ver detalhes do modelo no Apêndice B.1) e *simples-CIFAR* (Ver detalhes do modelo no Apêndice B.2) – baseados nas arquiteturas das redes *LeNet* (Lecun et al., 1998) e *Deep Belief Network* (Krizhevsky & Hinton, 2010) para a classificação de imagens nos conjuntos de dados MNIST e CIFAR, respectivamente. Os experimentos no banco de dados ImageNet foram realizados usando os modelos VGG16 (Simonyan & Zisserman, 2014) e ResNet50 (He et al., 2016) porque são os modelos base selecionados para a construção do modelo de descrição semântica global proposto.

5.6.2 Construção e Visualização do protótipo

A construção dos protótipo semânticos foi realizada assumindo como características das imagens de objetos aquelas características que foram extraídas da última camada densa (antes da camada *softmax*) dos modelos CNNs de classificação. Observa-se que para construir adequadamente o protótipo semântico proposto, a abordagem proposta

precisa de *elementos típicos* das categorias, ou qualquer informação sobre o valor de tipicidade (ou grau de representatividade) de um objeto para uma categoria específica.

A informação de tipicidade do objeto permite conhecer a priori a estrutura semântica interna (*organização prototípica*) da categoria. Ou seja, conhecer o *ground truth* da relevância visual de cada elemento da categoria permite identificar os objetos típicos necessários para calcular o protótipo semântico proposto. Aliás, o valor de tipicidade de cada elemento da categoria permite validar estatisticamente o modelo CPM proposto. Não obstante, nenhum dos bancos de dados utilizados nos experimentos possuem essa informação; de fato, não foram detectados atualmente bancos de dados de imagens rotuladas que possuam, além do rótulo da categoria, o valor de tipicidade de cada membro das categorias.

Como os bancos de dados de imagens usados não possuem rótulos com a informação de tipicidade dos objetos, não foi viável processar os protótipos usando essa informação proveniente dos banco de dados. Lake et al. (2015) mostraram que a saída da última camada dos modelos CNN pode ser usada como um sinal da aparência típica de uma imagem de entrada. Por esse motivo, e fundamentado nos estudos de Lake et al. (2015), assumiu-se nesta pesquisa como *membros típicos* da categoria apenas os elementos que são "inequivocamente" classificados como membros da categoria (*pontuação de tipicidade* (provabilidade)> 0.99) pelos modelos CNN de classificação. Finalmente, para cada categoria nos bancos de imagens utilizados, foram extraídas as características dos membros típicos e calculados os protótipos semânticos correspondentes (Vide Definição 1) usando o Algoritmo 1 visualizado na Figura 5.2. Nos experimentos realizados os protótipos foram calculados usando os modelos de classificação simples-MNIST (Ver detalhes do modelo no Apêndice B.1), VGG16 e ResNet50.

A Figura 5.4 mostra visualmente as representações dos protótipos semânticos calculados para as primeiras categorias do banco de dados MNIST (Ver Apêndice C.1 para mais exemplos). Apresenta-se, para cada representação dos protótipos semânticos (características de 128 dimensões), o intervalo dos valores esperados de cada característica unitária da categoria. Mostram-se também os valores de relevância de cada característica (Ω_i, b_i) aprendidos pelo modelo CNN de classificação usado (simples-MNIST).

Observa-se como, para cada uma das 128 características que compõem os protótipos calculados com o modelo simples-MNIST, existem variações próprias dos valores máximos e mínimos do valor médio das características e seus intervalos de valores esperados correspondentes. Essas variações das características dos membros da categoria são "particulares" de cada categoria. Dessa maneira, os protótipos semânticos calculados podem ser interpretados como a distribuição dos valores (ou sequência de DNA) das características típicas de cada categoria existente no banco de dados de

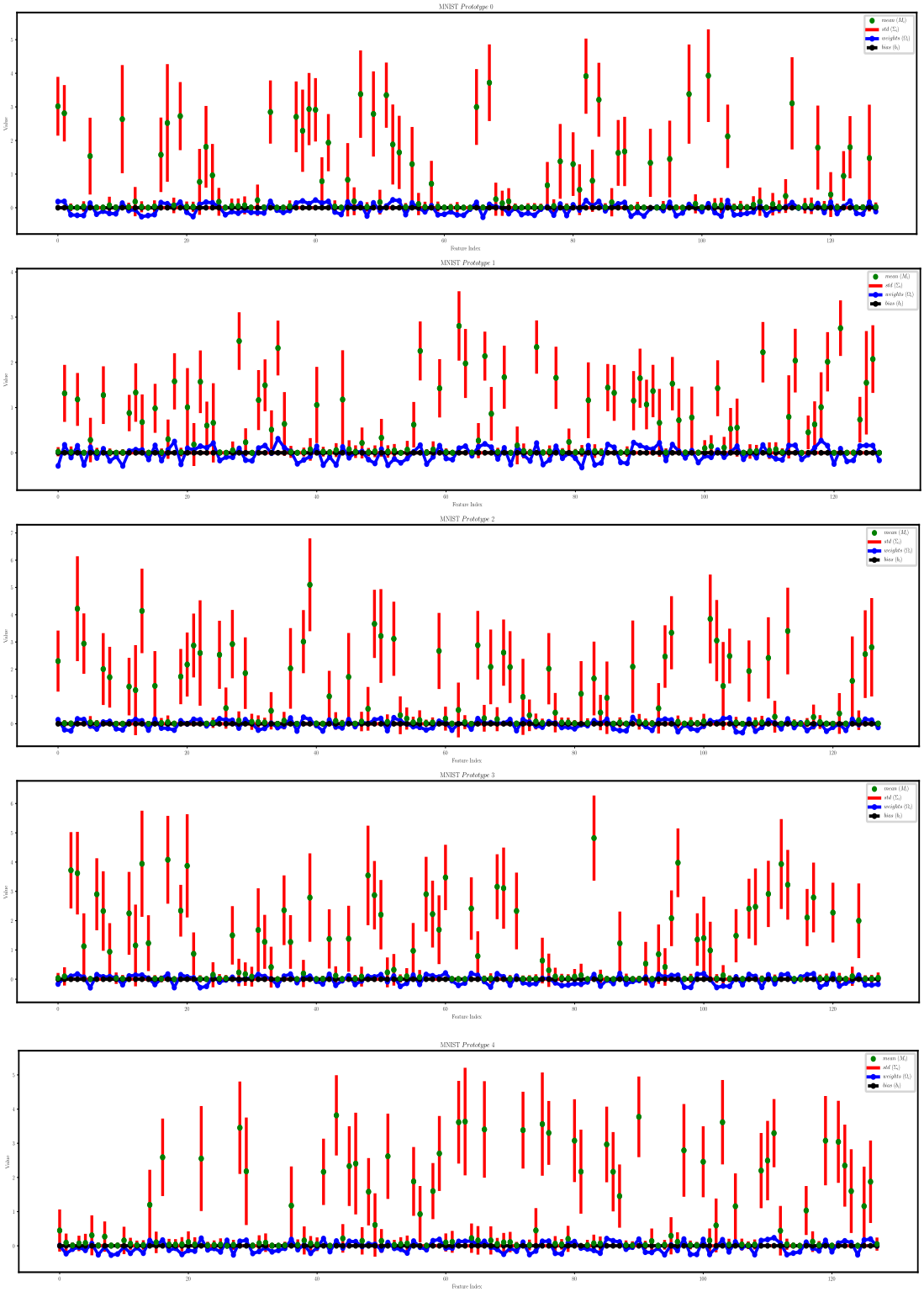


Figura 5.4: *Visualização da representação dos protótipos semânticos calculados para as 5 primeiras categorias do banco de dados MNIST.* Apresenta-se em cada representação, o intervalo dos valores das características unitárias da categoria, além dos valores de relevância correspondentes.

imagens. A grande dimensionalidade dos protótipos semânticos calculados no banco de dados ImageNet usando os modelo VGG16 e ResNet50, impede que os protótipos sejam visivelmente legíveis para serem mostrados nos resultados. O Apêndice C.3 mostra alguns exemplos visuais dos protótipos semânticos calculados para algumas categorias do banco de dados ImageNet usando o modelo VGG16.

5.6.3 Comportamento Prototípico

As características dos objetos e os valores de relevância das características unitárias da categoria permitem o agrupamento dos elementos em semelhança familiar dentro da categoria (Ver exemplos das Figuras 3.1 e 3.3). Aliás, em categorias prototípicas a representação da estrutura semântica interna da categoria é baseada na representatividade (tipicidade) dos membros que a compõem; e permite que objetos visualmente típicos sejam agrupados no centro semântico da categoria (*efeitos prototípicos*). A abordagem prototípica procura colocar os objetos típicos próximos ao centro semântico da categoria (próximos ao *protótipo abstrato*) e os elementos menos representativos mais distantes. O modelo CPM proposto procura reproduzir esse comportamento prototípico nas categorias de imagens de objetos baseando-se na representação do protótipo semântico e a distância prototípica propostas.

Contudo, não existe uma métrica definida para quantificar se a representação do protótipo semântico proposto captura corretamente o significado semântico central da categoria. O anterior é consequência de que não existe uma métrica definida para avaliar de forma robusta o nível de tipicidade de um objeto para uma categoria específica, uma habilidade reservada apenas para os seres humanos. A inexistência de bancos de dados rotulados que incluam essas informações de tipicidade (ou grau de representatividade visual) do objeto para cada categoria, não permite o uso de métodos estatísticos para avaliar de forma robusta o poder de explicação semântica do modelo CPM proposto. Consequentemente, são apresentadas outras abordagens para analisar a semântica por trás do modelo CPM proposto.

O protótipo abstrato construído para uma categoria específica (Ver Definição 2) constitui o centroide da categoria e representa o *centro semântico abstrato* da categoria. A *distância prototípica* proposta é calculada com relação a esse *centro semântico abstrato* da categoria. A distância semântica proposta visa mostrar que os elementos próximos do protótipo abstrato constituem elementos representativos ou típicos da categoria; e que os elementos distantes do protótipo abstrato representam os membros menos representativos da categoria, mas que possuem características que pertencem à categoria.

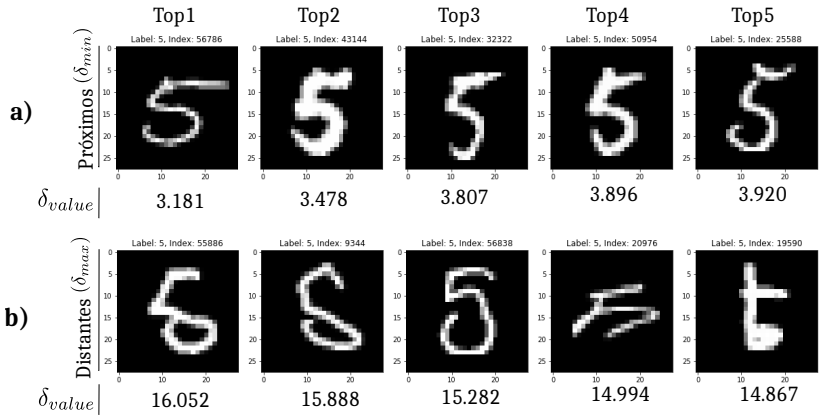


Figura 5.5: *Exemplos do comportamento prototípico capturado pelo modelo CPM proposto na categoria c_5 (número cinco) do banco de dados MNIST. a)* mostra-se, da esquerda a direita, o Top-5 dos elementos mais próximos do *protótipo abstrato* P_5 da categoria; *b)* são apresentados os Top-5 elementos mais distantes do *protótipo abstrato* da categoria. Para cada caso é apresentado o valor da distância prototípica (δ) e a posição (*índice*) de cada elemento dentro da categoria c_5 no banco de dados MNIST.

A Figura 5.5 constitui um exemplo visual que mostra a interpretação dos elementos da categoria "número cinco" usando o modelo CPM proposto. A Figura 5.5 apresenta o Top-5 dos membros identificados pelo modelo CPM proposto como os mais relevantes –visualmente– da categoria c_5 no banco de dados MNIST. O valor de relevância de cada membro da categoria foi calculado baseado na distância ao centro semântico da categoria (distância prototípica). Os Top-n membros mais relevantes da categoria são os n elementos mais próximos ao protótipo abstrato da categoria. Os Top-n membros mais distantes são os n elementos mais distantes ao protótipo. De maneira geral, os resultados obtidos para o Top-n elementos mais próximos e distantes do protótipo abstrato permitem assumir que quando é menor a *distância prototípica* (δ) do elemento, é maior a relevância visual (ou tipicidade) da imagem do elemento no contexto da categoria a qual pertence.

Conforme apresentado na Figura 5.5 a), o modelo CPM proposto identifica como elementos típicos da categoria c_5 (Top-5 mais próximos) os dígitos manuscritos que possuem características que são, sem dúvida, distintivas da categoria c_5 (do dígito cinco - 5). Como é apresentado nesse exemplo, o modelo consegue agrupar próximo do protótipo semântico P_5 os números cinco (5) bem formados ou com características típicas reconhecíveis pelos seres humanos. O modelo CPM proposto também pode identificar o significado semântico periférico da categoria (membros periféricos). Os elementos com características menos representativas da categoria c_5, ou pouco legíveis, são posicionados na periferia da categoria (Top-5 mais distantes). Esses elementos estão

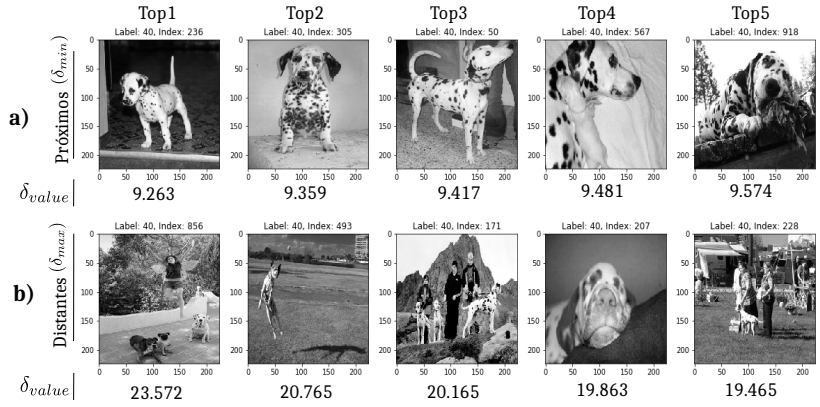


Figura 5.6: *Exemplos do comportamento prototípico capturado pelo modelo CPM proposto na categoria c_{40} (n02110341-dalmatian) do banco de dados ImageNet usando o modelo VGG16. a)* Top-5 dos elementos mais próximos do *protótipo abstrato* P_{40}. *b)* Top-5 dos elementos mais distantes do *protótipo abstrato* da categoria c_{40}-*dalmatian*.

posicionados longe do protótipo abstrato da categoria, mas dentro da categoria, porque ainda possuem características representativas da categoria. O exemplo apresentado na Figura 5.5 b) mostra que o modelo CPM (similar ao ser humano) identifica que esses elementos podem ser categorizados como o número 5, mas não são um número 5 típico.

A Figura 5.6 apresenta outro exemplo que mostra a interpretação semântica –da informação visual da imagem– realizada pelo modelo CPM proposto, mas nessa vez usando o modelo VGG16 como extrator de características no banco de dados ImageNet. De maneira similar ao exemplo anterior, a Figura 5.6 apresenta o Top-5 dos membros mais relevantes da categoria c_{40} (n02110341-*dalmatian*) com relação ao protótipo semântico P_{40} calculado no banco de dados ImageNet. Note-se que os membros reconhecidos pelo modelo CPM proposto como os mais próximos (semanticamente) ao protótipo da categoria (Figura 5.6 a)) são fáceis de reconhecer pelo ser humano, já que exibem as características típicas da categoria *dalmatian*. Observa-se também que os membros detectados pelo modelo como os mais distantes do protótipo (Figura 5.6 b)) –ou menos representativos da categoria *dalmatian*– apesar de conservarem algumas características da categoria, não são reconhecidos facilmente pelo ser humano. Ou seja, o modelo CPM proposto consegue identificar o Top-5 dos elementos mais/menos visualmente representativos da categoria, e – de maneira correta – reconhece como membros da periferia da categoria aqueles elementos onde não são identificadas todas as partes da categoria *dalmatian*; as cores típicas da categoria não são facilmente distinguíveis; ou, o tamanho, a forma e a pose dos membros não exibem essas características típicas da categoria, etc.

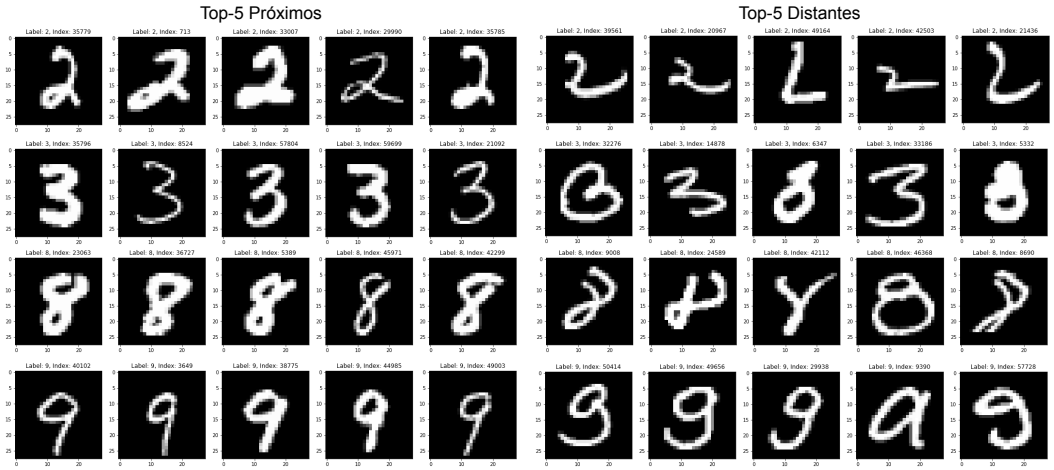


Figura 5.7: *Exemplos do comportamento prototípico capturado pelo modelo CPM proposto para uma amostra das categorias do banco de dados MNIST.*

As Figuras 5.7 e 5.8 apresentam outros exemplos dos Top-5 membros mais próximos e mais distantes –ao protótipo correspondente– para uma amostra de categorias dos bancos de dados MNIST e ImageNet, respectivamente (Ver mais exemplos no Apêndice D). Os resultados apresentados mostram que a métrica de distância semântica proposta (δ) pode agrupar os elementos relevantes da categoria (ou que possuem características típicas) próximos ao protótipo semântico da categoria. Também consegue agrupar os membros da categoria que não possuem características representativas bem distantes do centro semântico da categoria. Os resultados obtidos evidenciam que para cada categoria, os valores da distância mínima (δ_{min}) e da distância máxima (δ_{max}) ao protótipo correspondente são característicos de cada categoria.

Com base nos experimentos realizados, assumiu-se que o *protótipo semântico* proposto captura corretamente o *significado semântico central* da categoria. A *distância prototípica* definida, em conjunto com a representação semântica do protótipo, mostra ter influência na disposição dos elementos em torno do protótipo semântico da categoria, baseado na representatividade visual da imagem do objeto. Os resultados mostraram que os Top-5 dos elementos mais próximos ao protótipo semântico constituem membros representativos (ou típicos) da categoria. Ou seja, o modelo CPM proposto consegue identificar os Top-5 dos elementos mais próximos e mais distantes da categoria e os organiza prototipicamente dentro da categoria. Os objetos típicos da categoria são posicionados próximos ao protótipo abstrato e os menos característicos são posicionados mais distantes do centro semântico. Mas, o modelo CPM proposto consegue organizar todos os elementos da categoria com essa *organização prototípica*?

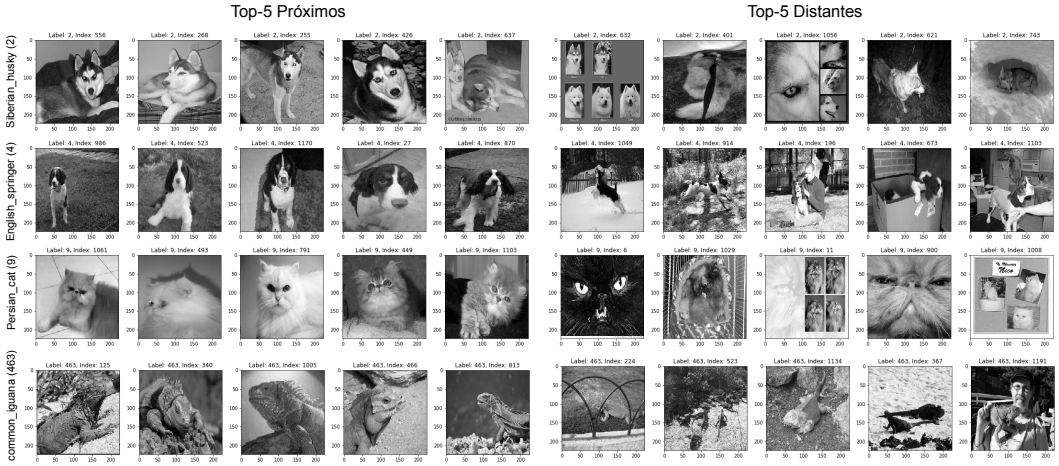


Figura 5.8: *Exemplos do comportamento prototípico capturado pelo modelo CPM proposto para uma amostra das categorias do banco de dados ImageNet usando o modelo VGG16.*

5.6.4 Organização prototípica da categoria

Conseguir que o modelo CPM proposto, baseado na interpretação da informação visual dos objetos, posicione semanticamente os objetos dentro da categoria simulando a organização prototípica da estrutura interna da categoria, constitui uma das motivações desta parte da pesquisa. Nesta secção apresentam-se alguns exemplos dos resultados alcançados com a metodologia desenvolvida para visualizar a estrutura interna das categorias.

Na Seção 5.5 apresentou-se a metodologia proposta para visualizar a estrutura semântica interna de uma categoria específica sem descartar características dos objetos. Foi proposta uma *função* $\rho : (F_{c_i}, \delta) \rightarrow (\mathbb{R}^2, L_1)$ que realiza um mapeamento (ou projeção) entre o espaço métrico das características m-dimensionais (F_{c_i}, δ) e o espaço cartesiano usando a métrica L1 $(\mathbb{R}^2, L_1)$. As relações existentes entre as métricas de distâncias de ambos os espaços métricos (Ver Proposição 1 e Apêndice G para exemplos) converte à projeção ρ em uma *função continua* entre ambos os espaços métricos.

Organização Prototípica no domínio $\mathbb{R}^2$

Consequentemente, visualizar a estrutura interna de uma categoria específica consiste em visualizar, no espaço métrico $(\mathbb{R}^2, L_1)$, cada elemento da i-ésima categoria $c_i \in C$ mapeado através da função ρ. Pelas propriedades da função ρ, se a i-ésima categoria mapeada $\rho(c_i)$ for organizada em torno do protótipo mapeado $\rho(P_i)$ baseado na representatividade visual dos elementos dentro da i-ésima categoria, então a representação

semântica proposta para a categoria consegue simular a organização prototípica dos membros da categoria.

Por conseguinte, todo comportamento observado em termos da métrica de distância L_1 no espaço métrico $(\mathbb{R}^2, L_1)$, é um comportamento que preserva-se em termos da métrica de distância δ no espaço métrico (F_{c_i}, δ). Preservar a posição semântica no espaço métrico $(\mathbb{R}^2, L_1)$ significa que os objetos típicos no espaço métrico (F_{c_i}, δ) continuam sendo encontrados próximos do protótipo mapeado $\rho(P_i)$, e que os menos representativos –quando for mapeados– também permanecem distantes do centro semântico da categoria mapeada.

A função ρ mapeia membros do espaço métrico (F_{c_i}, δ) para um ponto (x, y) no espaço métrico $(\mathbb{R}^2, L_1)$ onde x representa o *valor semântico* do objeto e y representa a sua *distância prototípica*. Observa-se como, pela propriedade de *identidade de indiscernível* das métricas de distância de ambos os espaços métricos, a distância do protótipo (ou do protótipo mapeado $\rho(P_i)$) a si mesmo é zero. Ou seja, nos experimentos apresentados nesta seção, o protótipo mapeado $\rho(P_i)$ sempre aparece com valor zero no eixo da *distância prototípica* (eixo y). Também, como a distância prototípica é *não-negativa* por natureza, os elementos mais semelhantes ao protótipo de uma categoria específica são aqueles que são posicionados próximos ao valor zero (o protótipo) no eixo y. Aliás, a Definição 5.5 define que o *valor resumo* do protótipo semântico da

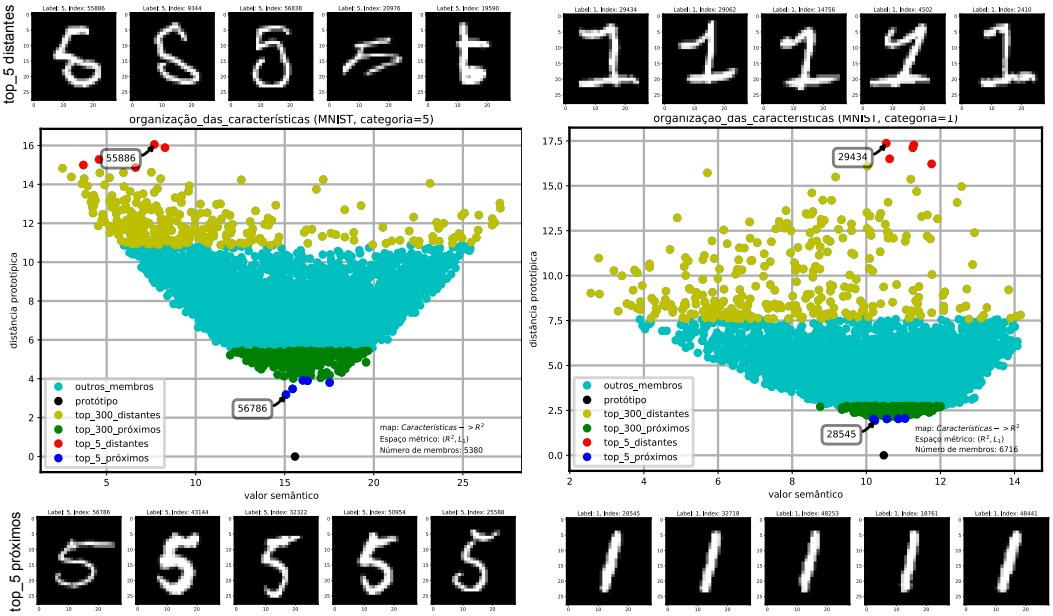


Figura 5.9: *Organização prototípica dentro das categorias* 5 *e* 1 *no banco de dados MNIST.* Na parte superior e inferior, apresentam-se os Top-5 elementos mais distantes e mais próximos do protótipo semântico P_5 e P_1 respectivamente. O índice do elemento mais próximo e do mais distante é anotado dentro da caixa preta.

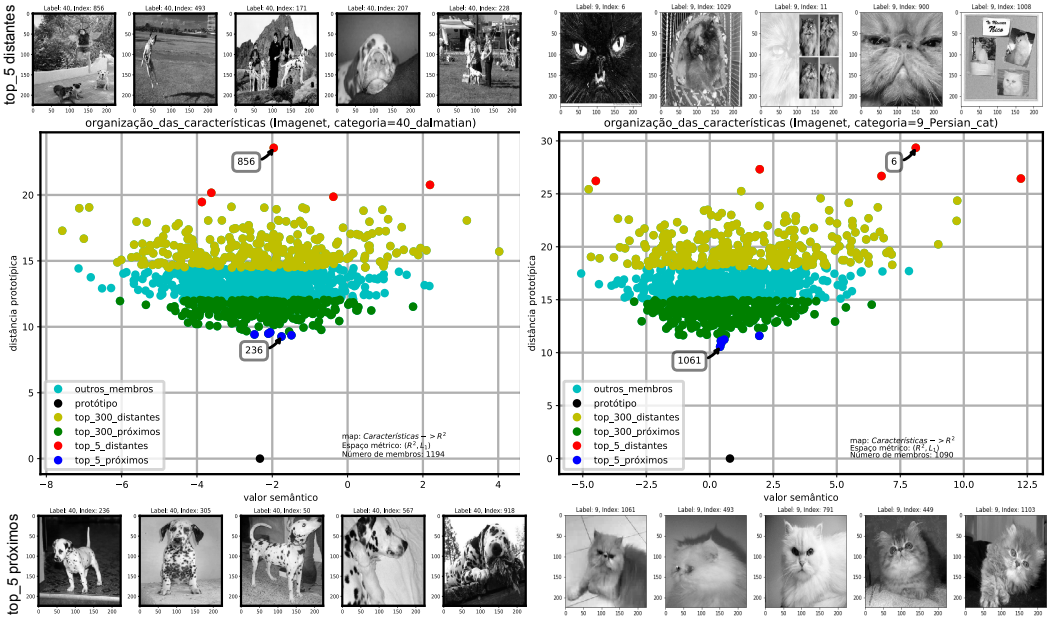


Figura 5.10: *Organização prototípica dentro das categorias c_{40}-dalmatian e c_9-Persian cat do banco de dados ImageNet para características extraídas com o modelo VGG16.* Na parte superior e inferior, apresentam-se os 5 elementos mais distantes e mais próximos do protótipo semântico P_{40} e P_9 respectivamente. O índice do elemento mais próximo e do mais distante é anotado dentro da caixa preta.

categoria constitui o *valor semântico da categoria*. Em torno desse valor semântico da categoria (no eixo x) serão encontrados todos os valores semânticos dos membros da categoria.

A Figura 5.9 mostra a representação –usando a abordagem proposta– da estrutura semântica interna das categoria c_5 (esquerda) e c_1 (direita) do banco de dados MNIST. Apresenta-se o espaço m-dimensional das características dessas categorias –(F_{c_5}) e (F_{c_1})– mapeadas no espaço métrico $(\mathbb{R}^2, L_1)$ usando a função contínua ρ. Observa-se como os Top-5 elementos mais próximos/distantes da categoria no espaço métrico (F_{c_i}, δ) quando são mapeados preservam a posição semântica no domínio $\mathbb{R}^2$. Ou seja, os 5 elementos mais próximos (pontos em azul) e os mais distantes (pontos em vermelho) dos protótipos semânticos P_5 e P_1, continuam sendo os mesmos no espaço métrico $(\mathbb{R}^2, L_1)$ que os calculados no espaço métrico (F_{c_i}, δ).

Para cada uma das categorias apresentadas na Figura 5.9, o tamanho da amostra usada corresponde com a quantidade dos elementos corretamente classificados pelo modelo CNN usado; ou seja, os membros da categoria. A visualização dos Top-300 elementos mais próximos e dos Top-300 elementos mais distantes em diferentes cores permite perceber uma organização prototípica na estrutura interna das categorias c_5 e c_1 do banco de dados MNIST. Por exemplo, se pode perceber que os objetos típicos da

categoria c_5 (Ver Figura 5.5a)) são posicionados semanticamente próximos do protótipo mapeado $\rho(P_5)$ e os objetos menos representativos (Ver Figura 5.5b)) são encontrados distantes do centro semântico da categoria.

A Figura 5.10 mostra dois exemplos da organização prototípica alcançada pela abordagem proposta na estrutura interna das categorias c_{40}-*dalmatian* e c_9-*Persian cat* do banco de dados ImageNet. Observa-se como, de maneira semelhante ao exemplo da Figura 5.9, os elementos da categoria mapeados preservam a posição semântica no domínio $\mathbb{R}^2$ em torno do protótipo correspondente. Assim, os Top-5 elementos mais próximos/distantes da categoria no espaço métrico (F_{c_i}, δ) preservam sua posição semântica no domínio $\mathbb{R}^2$.

Observe que as particularidades de como o mapeamento é realizado usando a *função continua* $\rho : (F_{c_i}, \delta) \rightarrow (\mathbb{R}^2, L_1)$ permitem realizar uma análise da representação da categoria no espaço métrico $(\mathbb{R}^2, L_1)$. Consequentemente, os fenômenos semânticos da representação da categoria observados no espaço métrico $(\mathbb{R}^2, l_1)$ são extensivos ao espaço m-dimensional das características dos objetos da categoria (F_{c_i}, δ).

5.6.5 Captura da Tipicidade Visual

Nesta seção, são analisadas as possíveis relações existentes entre o *valor semântico*, a *distância prototípica* e a *tipicidade visual* do objeto. A inexistência de bancos de dados com anotações de pontuação de tipicidade de imagens impossibilita a avaliação robusta das relações existentes entre o *valor semântico*, a *distância prototípica* e a *tipicidade visual* do objeto. Como alternativa, foram usadas outras abordagens para analisar a semântica por trás da representação proposta e observar como pode influenciar as variações dessas variáveis (valor semântico e distância prototípica) na informação visual do objeto (tipicidade).

O valor semântico e a distância prototípica

Os resultados alcançados na análise realizada da estrutura interna das categorias –no espaço métrico $(\mathbb{R}^2, L_1)$– permitiu observar que a *forma* da estrutura interna da categoria mapeada depende fortemente do modelo CNN usado e das características do banco de imagens. Obviamente, o anterior é consequência de que os valores das variáveis *valor semântico* e *distância prototípica* estão influenciados pelas características (arquitetura, pesos aprendidos) do modelo CNN usado para a extração de características. Ou seja, o *valor semântico* (significado semântico do objeto) e a *distância prototípica* definem como é agrupada a categoria; e transitivamente, o modelo CNN usado define como é interpretada e estruturada a categoria.

Precisamente, a maneira de como são aprendidos os pesos do modelo CNN usado, define a relação existente entre os valores do *valor semântico* e da *distância prototípica*. Por definição, o valor da distância prototípica estará fortemente correlacionado com o valor semântico, sempre que seja garantido que os pesos aprendidos pelo modelo CNN na camada *softmax* sejam estritamente positivos ($\omega_{ij} \geq 0$, $\forall \omega_{ij} \in \Omega_i$)(Ver Seção 5.4.1). Mas, os modelos CNN pré-treinados usados não cumprem com esse pré-requisito dos pesos aprendidos na camada softmax. Os experimentos realizados nos conjuntos de dados MNIST, CIFAR e ImageNet –com os modelo CNN correspondentes– mostram que existe uma pequena força de associação linear entre essas duas variáveis (coeficiente de *Pearson* com valores entre [-0.4, -0.3] e [0.3, 0.4]), mas não permite concluir que é possível generalizar um padrão de comportamento entre o valor semântico e a distância prototípica do objeto.

Pontuação de Tipicidade

Lake et al. (2015) mostraram que a força da resposta de classificação da última camada dos modelos CNN para a categoria de interesse, pode ser usada como um sinal de quão típica é a imagem de entrada. O valor de tipicidade usado nos experimentos de Lake et al. (2015) constitui o valor semântico do objeto definido nesta pesquisa (Ver Definição 9).

Em contraste com os resultados de Lake et al. (2015), os experimentos realizados com os modelos VGG16 e ResNet50 no banco de dados ImageNet mostraram que usar o valor semântico como pontuação de tipicidade dos objetos pode ser problemático, pois imagens de objetos com igual valor semântico não implica que os objetos sejam igual de representativos (visualmente típicos) dentro da categoria. A Figura 5.10 (direita, categoria c_9-*Persian cat*) mostra um exemplo desse resultado. Observa-se como na categoria *Persian cat* o quinto elemento do Top-5 elementos mais próximos do protótipo tem um valor semântico semelhante com o elemento da posição do Top-5 mais distante (aqueles com valor semântico ≈ 2), mas esses elementos são visualmente muito diferentes. Ou seja, representar a tipicidade da imagem do objeto com o valor semântico poderia ser uma condição necessária, mas não é uma condição suficiente. Por outro lado, observa-se como a distância prototípica proposta sim consegue capturar a diferença de tipicidade visual entre essas duas imagens de objetos e identificar qual delas é mais representativa da categoria *Persian cat*.

A Figura 5.10 (esquerda, categoria c_{40}-dalmatian) também mostra outro exemplo do problema de usar o valor semântico como pontuação da tipicidade do objeto. Nota-se como no elemento anotado com índice 856 não se encontram características

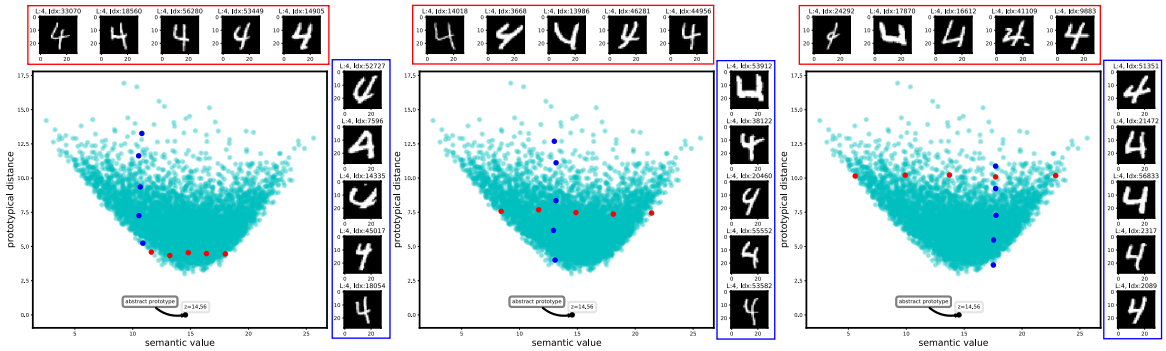


Figura 5.11: *Análise do comportamento da tipicidade dentro da categoria c_4 do banco de dados MNIST.* Apresentam-se exemplos da variação de aparência visual de membros da categoria quando o valor semântico permanece constante e varia a distância prototípica (em azul); e exemplos onde varia o valor semântico e a distância prototípica permanece constante (em vermelho).

representativas da categoria *dalmatiam*, mas o elemento possui um valor semântico aproximado ao valor semântico esperado da categoria (valor semântico do protótipo) e ao valor semântico de outros elementos visualmente típicos da categoria (exemplo o elemento anotado com índice 235). Aliás, observa-se como o modelo CPM proposto sim consegue captar melhor essa diferença de tipicidade visual entre esses elementos e colocar elementos visualmente representativos mais próximos ao protótipo (Ver *Top_5* próximos/distantes).

Para realizar uma análise mais detalhada do impacto do *valor semântico* vs *distância prototípica* na representatividade visual do objeto foram realizados outros experimentos que visam observar qual é o comportamento da informação visual da imagem quando uma dessas variáveis semânticas mantém seu valor constante enquanto a outra muda ascendentemente de valor. As Figuras 5.11 e 5.12 mostram exemplos desse experimento para as categorias c_4 e c_9 dos bancos de dados MNIST e ImageNet, respetivamente (Ver Apêndice E para outros exemplos). Ambas as Figuras apresentam exemplos da variação de aparência visual de membros da categoria quando o *valor semântico* do elemento permanece constante e quando o valor da *distância prototípica* varia ascendentemente (ver membros anotados em azul). Também são apresentados exemplos do experimento oposto, onde o valor semântico dos elementos aumenta de maneira ascendente o a distância prototípica desses elementos é praticamente a mesma (ver membros anotados em vermelho).

Os exemplos apresentados demonstram – por contraexemplo – que, contrário à conclusão de Lake et al. (2015), o valor semântico do objeto não consegue capturar de maneira totalmente robusta a tipicidade da imagem do objeto nas categorias de

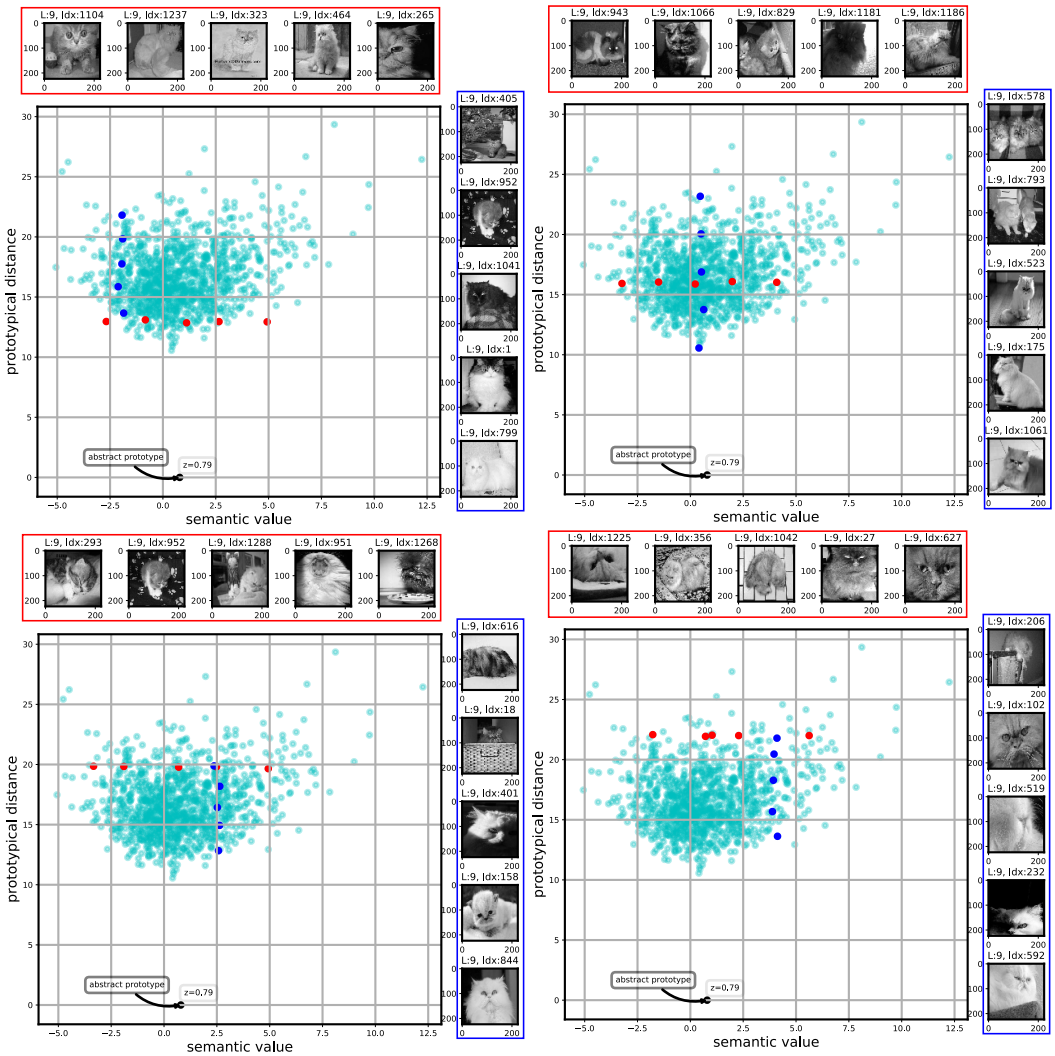


Figura 5.12: *Análise da tipicidade dentro da categoria c_9-Persian cat do banco de dados ImageNet para características extraídas com o modelo VGG16.* Apresentam-se exemplos da variação de aparência visual de membros da categoria quando o valor semântico permanece constante e varia a distância prototípica (em azul); e exemplos onde varia o valor semântico e a distância prototípica permanece constante (em vermelho).

análise. Basta observar os membros anotados em azul nas Figuras 5.11 e 5.12 para apreciar que elementos com igual valor semântico podem ser totalmente diferentes visualmente. Aliás, observa-se como para uma distância prototípica fixa (ver elementos anotados em vermelho) a variação ascendente do valor semântico do objeto não gera mudanças significativas na relevância visual da imagem como para reconhecer um patrão de comportamento; ou seja, uma diminuição ou aumento do valor semântico não significa menos ou maior relevância visual da imagem do objeto para a categoria.

Por outro lado, os resultados apresentados mostram que a distância prototípica

proposta, comparada com o valor semântico, consegue capturar melhor a tipicidade visual do objeto. Observa-se nas Figuras 5.11 e 5.12 que para um valor semântico fixo (ver elementos anotados em azul), a variação ascendente da distância prototípica gera mudanças na tipicidade da informação visual da imagem ordenadas de maneira decrescente; ou seja, para um valor semântico fixo, um aumento da distância prototípica sempre implica menos representatividade da imagem na categoria de análise. Nota-se também que elementos com igual distância semântica (anotados em vermelho) possuem uma aparência visual semelhante. Além disso, percebe-se que esses membros da categoria (anotados em vermelho) possuem uma relevância de tipicidade visual que está em concordância com quão próximos ou distantes estão do protótipo abstrato da categoria.

Os experimentos realizados permitem assumir que o modelo CPM proposto consegue capturar (interpretar) o significado semântico central das categorias de objetos. Os resultados mostraram que, mesmo com diferentes modelos CNN de extração de características e conjuntos de imagens, o modelo CPM proposto consegue organizar os membros dentro da estrutura interna da categoria seguindo uma organização prototípica.

Aliás, os experimentos também mostraram que assumir como "pontuação de tipicidade" do objeto a *distância prototípica* proposta, mostra ser uma abordagem mais robusta que basear-se somente no *valor semântico* do objeto como pontuação de tipicidade da imagem. São vários as rações que sustentam a afirmação anterior:

i) o valor semântico dos elementos de uma categoria pode ser positivo ou negativo (Ver exemplo da Figura 5.12), e mesmo uma mudança de signo no valor semântico do objeto pode que não gere uma mudança significativa na tipicidade visual da imagem correspondente. Contrariamente, a *distância prototípica* proposta é não negativa por definição;

ii) a variação de *uma unidade positiva* do valor semântico do elemento é ambígua, pois essa variação não significa um aumento ou diminuição da relevância visual da imagem na categoria. Por outro lado, uma variação positiva no valor da distância prototípica proposta significa menos tipicidade do elemento;

iii) definir como estritamente positiva a importância das características (Ω_i) para o cálculo da distância prototípica impede que na soma do produto escalar alguns términos sejam anulados. Esse fenômeno acontece no cálculo do valor semântico do elemento (Ver Equação 9), e como consequência, elementos bem diferentes semanticamente podem ter o mesmo valor semântico;

iv) os resultados mostram que a *distância prototípica* proposta é inversamente proporcional à relevância visual do objeto dentro da categoria. Ou seja, a métrica de distância prototípica proposta pode ser entendida como uma pontuação de tipicidade do objeto dentro da categoria (*pontuação de tipicidade* $(o) = 1/\delta(o, P_i)$). Aliás, o comportamento observado do *valor semântico* do objeto dentro da categoria não permite generalizar sua relação com a tipicidade do objeto;

v) a distância prototípica proposta pode entender-se como a diferença do significado semântico do objeto e do significado semântico do protótipo da categoria. Ou seja, a métrica de tipicidade proposta inclui em si mesma o valor semântico do objeto;

vi) usar uma pontuação de tipicidade do objeto que representa uma métrica no domínio das características m-dimensionais do objeto, permite analisar fenômenos semânticos e transformações no espaço m-dimensional de maneira parecida a como são analisadas no espaço Euclidiano.

5.7 Discussão

No presente capítulo foi proposto o modelo CPM que visa representar o significado semântico central das categoria de objetos: o protótipo. A modelagem do *significado semântico central* foi fundamentada nos estudos da Semântica Cognitiva relacionados com a *Teoria dos Protótipos*. Aliás, foi apresentado um algoritmo simples para construir, representar e armazenar o protótipo semântico das categorias de objetos usando qualquer modelo CNN de classificação. O método proposto transfere o conhecimento dos modelos CNN de classificação pré-treinados para uma estrutura semântica (protótipo semântico) que visa representar o significado semântico central das categorias aprendidas.

Nos experimentos realizados constatou-se que o modelo CPM proposto (protótipo semântico e distância semântica) permite a simulação da organização prototípica na estrutura interna das categorias aprendidas. Os experimentos também mostraram que a métrica de distância semântica proposta, no domínio das características do objeto, pode ser entendida como uma pontuação de tipicidade do objeto para uma categoria específica. A métrica de distância semântica proposta constitui uma generalização das distâncias semânticas proposta pelos modelos psicológicos formais MPM e GCM. Consequentemente, o modelo CPM proposto permite trazer à luz os estudos de Ciências Cognitivas relacionados com a Teoria dos Protótipos às Redes Neuronais Convolucionais. No próximo capítulo apresenta-se como usar o modelo CPM proposto na descrição semântica global de objetos.

Capítulo 6

Descritor Semântico Global

A maioria dos métodos existentes na literatura para a descrição de características da imagem (Han et al., 2015; Lin et al., 2016b; Simo-Serra et al., 2015; Zagoruyko & Komodakis, 2015; Zbontar & LeCun, 2015; Choy et al., 2016; Han et al., 2017; Kim et al., 2017; Rocco et al., 2018) representam a informação semântica da característica usando um leque de abordagens diferentes. Mas, nenhum desses descritores constrói as assinaturas encapsulando a informação semântica sob o extenso embasamento teórico existente na Semântica Cognitiva para representar o significado.

Diferente dos modelos existentes na literatura, o presente trabalho propõe introduzir os fundamentos teóricos da semântica cognitiva relacionados com a Teoria dos Protótipos para representar o *significado semântico* da informação contida na imagem. A teoria proposta por Rosch (Rosch, 1975b; Rosch & Mervis, 1975) estabelece que o ser humano aprende o significado semântico das categorias (o protótipo) e o inclui nos processos cognitivos. Sob essas premissas, a ideia principal do modelo de descrição semântica proposto sugere usar o protótipo semântico da categoria como entidade semântica que rege a representação semântica dos componentes básicos (os objetos) que compõem o significado contido na imagem.

A abordagem de descrição semântica de objetos proposta nesta pesquisa pretende simular duas premissas principais da abordagem humana de descrição global de objetos: *i)* descrever os objetos baseado nas mesmas características aprendidas para categorizá-los; *ii)* descrever globalmente o objeto usando a estratégia de destacar aquelas características que o tornam diferente (distintivo) dentro da categoria à qual pertence. Observa-se que com essa abordagem os seres humanos são capazes de construir descrições como: *a)* o dálmata é um cão (categoria base) branco com manchas pretas (características que o distingue dentro da categoria cão); *b)* a orca distingue-se dos golfinhos (categoria base) por seu maior tamanho e pelo padrão de cores preto e

branco de sua pele; *c)* a zebra é um equino (categoria base) cuja característica mais distintiva é sua cor baseada em listras pretas e brancas. Mas, como modelar uma descrição global do objeto com um comportamento semelhante? Ou seja, reconhecida a categoria à qual o objeto pertence, como identificar quais são as características que o distinguem dentro dela?

O presente capítulo propõe um modelo de descrição semântica global de imagens de objetos que pretende simular esse comportamento humano em tarefas de descrição semântica de imagens. Especificamente, apresenta-se como introduzir a representação do protótipo semântico proposto no fluxo de processos da hipótese de descrição semântica de objetos apresentada na Figura 1.2.

6.1 O protótipo na descrição global do objeto

No Capítulo 5 foram apresentados resultados que mostram que o modelo CPM proposto consegue simular a *organização prototípica* dos elementos dentro da categoria. A abordagem proposta para analisar a estrutura semântica interna da categoria permitiu observar que objetos que compartilham características semelhantes podem ser encontrados nas mesmas posições semânticas dentro do espaço métrico da categoria (*family resemblance structure*). Observa-se que essa posição semântica dentro da categoria está relacionada com o significado semântico do objeto (o valor semântico) e com o valor de representatividade, capturado pelo modelo CPM proposto, desses elementos dentro da categoria (a distância prototípica).

Os experimentos realizados no Capítulo 5 mostraram que esses atributos do objeto (o valor semântico e a distância prototípica) conseguem posicionar o objeto em uma posição semântica *distintiva* dentro da categoria. O *valor semântico* pode ser entendido como o valor resumo do significado semântico do objeto usado pelos modelos CNN de classificação (Ver Definição 9); e a *distância prototípica* pode ser entendida como a pontuação de representatividade ou medida de quão diferente é o objeto com relação protótipo semântico da categoria.

A abordagem de descrição semântica de objetos baseada em protótipos proposta (ver Figura 1.2) pretende descrever o objeto usando seu *significado semântico* (processo de generalização) e sua *distinção semântica* dentro da categoria (processo de discriminação). Sob esses pressupostos, é proposto um *modelo de descrição semântica global de objetos baseado em protótipos* que constrói uma representação semântica do objeto usando seu *significado semântico* (valor semântico) e suas diferenças com relação ao significado semântico central da categoria (distância prototípica).

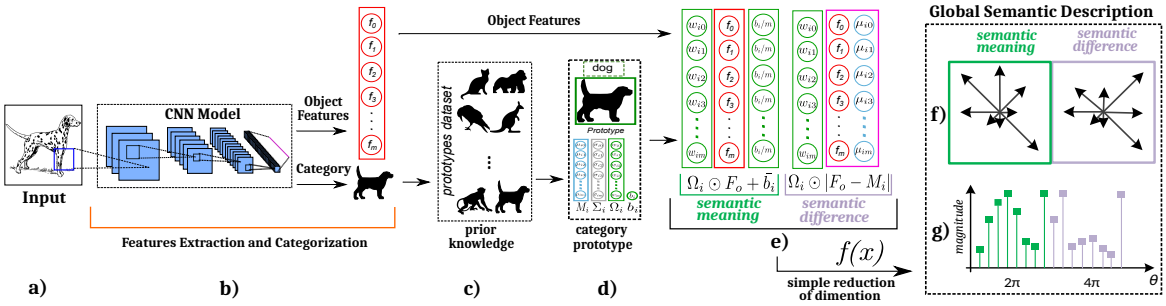


Figura 6.1: *Modelo de descrição semântica de objetos baseado em protótipos.* Fluxo de processos que transformam as características extraídas do objeto de entrada na assinatura do Descritor Semântico Global proposto. *a)* imagem de entrada; *b)* extração de características e categorização da imagem usando um modelo CNN; *c)* banco de protótipos; *d)* seleção do protótipo da categoria do objeto de entrada usando a saída softmax do modelo CNN; *e)* representação semântica do objeto usando o protótipo semântico; *f)* representação gráfica do *Descritor Semântico Global* resultante da função de redução de dimensionalidade ($f(x)$); e *g)* assinatura do Descritor Semântico Global proposto. Fonte: Elaborado pelo autor.

6.1.1 O vetor significado semântico

Nesta pesquisa assume-se como *significado semântico* do objeto aquele *significado da informação visual da imagem* usado pela camada softmax dos modelos CNN para classificar o objeto (Ver seção 5.4). Assim, a abordagem de descrição semântica global de objeto baseada em protótipos entende que o *vetor do significado semântico de objeto* é o vetor semântico ($\vec{z} = \Omega_i \odot F_o + \bar{b_i}$) construído a partir do *produto de Hadamard* (*element-wise operations*) entre os vetores usados para calcular o *valor semântico* do objeto (Ver Definição 9). O vetor da representação do significado semântico do objeto usa o vetor *bias* ($\bar{b_i}$) para dissolver uniformemente o *bias* aprendido para i-ésima categoria ($b_i = \sum_m \bar{b_i}$) em cada componente do vetor significado semântico ($\vec{z}$). Observa-se que com essa abordagem, basta uma soma de cada componente do vetor de significado semântico para recuperar o valor semântico do objeto ($z = \sum_m \vec{z}$). Consequentemente, o vetor do significado semântico do objeto proposto contém as mesmas definições semânticas usadas pelos modelos CNN para categorizar o objeto dentro de uma categoria específica.

6.1.2 O vetor diferença semântica

Aliás, assume-se que a distinção semântica de um objeto para uma categoria específica pode ser representada como a discrepância semântica entre as características do objeto e as características do elemento ideal (elemento mais prototípico) da i-ésima catego-

ria, o protótipo abstrato da i-ésima categoria. Como as características do objeto (F_o) e o protótipo abstrato da i-ésima categoria ($M_i \in P_i$) pertencem ao mesmo domínio m-dimensional das características (espaço métrico das características), a distância prototípica proposta pode ser usada como medida do caráter distintivo dos objetos dentro da categoria.

Consequentemente, a abordagem de descrição semântica proposta assume como o vetor *distinção semântica do objeto*, o vetor de *diferença semântica* ($\vec{\delta} = \Omega_i \odot |F_o - M_i|$) construído a partir do *produto de Hadamard* (*element-wise operations*) entre os vetores usados para calcular a distância prototípica do objeto (Ver Definição 5). Nota-se que o vetor *diferença semântica* do objeto (ou vetor distância prototípica) pode ser entendido como o *vetor semântico residual* ($\vec{r} = |F_o - M_i|$) ponderado com relevância relativa das características para a i-ésima categoria (Ω_i). Ou seja, o *vetor diferença semântica* ($\vec{\delta}$) do objeto está constituído pelos valores absolutos da diferença entre cada característica do objeto com cada característica do protótipo abstrato da categoria, esses valores encontram-se normalizados com os valores da relevância de cada característica unitária da categoria (Ver Definição 5).

Observa-se também que, quando o modelo CNN garanta pesos não negativos na camada softmax, o *vetor diferença semântica* do objeto pode ser entendido como a soma da diferença absoluta entre o vetor *significado semântico de objeto* e o vetor *significado semântico central* de i-ésima categoria (Ver seção 5.4.1). Assim, similar à representação do vetor significado semântico do objeto, a representação proposta para o *vetor diferença semântica* de objeto possui a vantagem de que a soma dos elementos do vetor é suficiente para recuperar a *distância prototípica* (ou pontuação de tipicidade) do objeto ($\delta = \sum_m \vec{\delta}$).

A Figura 6.1 mostra uma visão geral do *modelo de descrição semântica de objetos baseado em protótipos* proposto. A Figura 6.1 mostra o fluxo de processos que transformam a imagem de entrada na assinatura do *Descritor Semântico Global* (*Global Semantic Descriptor based on Prototypes (GSDP)*) proposto. Nota-se que o descritor GSDP proposto tem como pré-requisito o conhecimento prévio (Figura 6.1c) de cada protótipo semântico das categorias correspondentes ao modelo CNN usado. Esse banco de protótipos semânticos debe ser pré-calculados *off-line* usando o Algoritmo 1.

Na Figura 6.1, após dos processos de extração de características e categorização do objeto usando um modelo CNN de classificação (Figura 6.1b), usou-se o protótipo semântico da categoria correspondente para a construção da representação semântica proposta que visa descrever semanticamente as características do objeto. O representação mostrada na Figura 6.1e) apresenta visualmente como introduzir o protótipo semântico da categoria, através do vetor *significado semântico* ($\vec{z}$) e do vetor *diferença*

Algoritmo 2 Descritor Semântico Global ψ

1: ***Entrada***: Imagem de um objeto o
2: ***Saída***: Assinatura semântica do objeto (ψ_o)
3: ***Dados***: Modelo CNN Λ, *banco_de_protótipos*
4: $F_o, c_i \leftarrow \Lambda.características\_e\_predição(o)$
5: $M_i, \Sigma_i, \Omega_i, b_i \leftarrow banco\_de\_protótipos(c_i)$
6: $significado \leftarrow f\left(F_o, \Omega_i, b_i, significado\right)$
7: $diferença \leftarrow f\left(\left|F_o - M_i\right|, \Omega_i, b_i, diferença\right)$
8: **return** $significado \oplus diferença$

semântica ($\vec{\delta} = \Omega_i \odot \vec{r}$), na descrição semântica global de imagens de objetos.

Observa-se que os processos *b)*, *c)* e *d)* mostrados na Figura 6.1 correspondem aos processos relacionados com o reconhecimento do protótipo semântico que representa visualmente a imagem de entrada. Esse processo de reconhecimento e recuperação do protótipo é realizado, neste caso, categorizando o objeto com o modelo CNN de classificação e selecionando no banco de protótipos calculados o protótipo correspondente à categoria em questão. O Algoritmo 2 e a Figura 6.1 detalham os principais passos da abordagem de descrição semântica baseada em protótipos proposta. Observa-se que as etapas apresentadas seguem o mesmo fluxo de trabalho da hipóteses de descrição humana de objetos apresentada na Figura 1.2.

Uma desvantagem da representação semântica do objeto proposta na Figura 6.1e) é que possui uma alta dimensionalidade. A representação é construída baseada no vetor *significado semântico* ($\vec{z} = F_o \odot \Omega_i$) e no vetor *diferença semântica* ($\vec{\delta} = |F_o - M_i| \odot \Omega_i$), os quais são vetores m-dimensionais que possuem uma alta dimensionalidade. A grande dimensionalidade da representação semântica do objeto proposta torna seus usos práticos inviáveis em tarefas comuns de Visão Computacional (Han et al., 2017; Kim et al., 2017). Para lidar com essa desvantagem, foi construída uma função de redução de dimensionalidade $f(x)$ que visa comprimir a informação contida na representação semântica baseada em protótipos (Figura 6.1e)), em uma representação mais compacta e de menor dimensionalidade (Ver Figura 6.1 f) e g)) que preserva as particularidades da representação semântica original.

6.2 Redução da dimensionalidade

A maioria dos algoritmos existentes de redução de dimensionalidade como PCA (Abdi & Williams, 2010) e NMF (Lee & Seung, 2001), baseiam-se no descarte das características que não geram variação significativa. Embora essa abordagem possui um bom desempenho em várias tarefas, com a aplicação desses algoritmos perde-se a capacidade

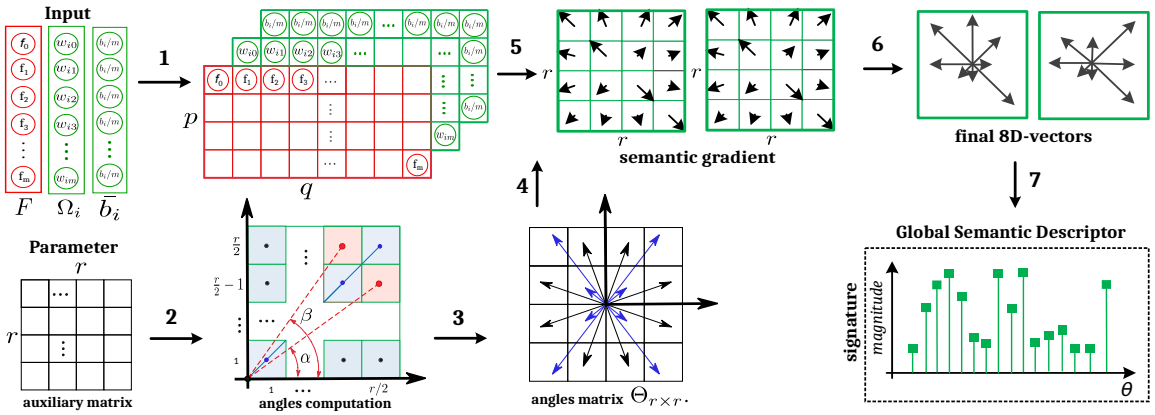


Figura 6.2: *Redução de dimensionalidade.* Funcionamento interno da função de transformação $f(x)$, que converte a representação semântica do objeto proposta na assinatura GSDP correspondente. A transformação $f(x)$ retorna uma assinatura final construída pela concatenação de cada 8D-vetor calculado a partir de cada gradiente semântico unitário. Mostra-se o caso trivial correspondente a vetores m-dimensionais de entrada correspondentes ao significado semântico do objeto e que possuem 2 vezes a dimensão da matriz auxiliar ($m = p{\cdot}q$ e $p = r$; $q = 2r$); consequentemente, a assinatura de saída (16D) está constituída por 2 vetores 8D. Fonte: Elaborado pelo autor.

de interpretação semântica dos dados (Abdi & Williams, 2010). Sob a perspectiva da Teoria dos Protótipos a abordagem de descartar características não é considerada adequada no espaço semântico, devido à ausência de definições de atributos necessários e suficientes para categorizar um objeto (*intensional non-discreteness*). As vezes, o fato de descartar características pode significar eliminar o poder discriminativo entre elementos da categoria; pois podem existir alguns objetos dentro da categoria que não possuem algumas das características típicas da categoria. Por exemplo, na categoria pássaro a característica voar constitui uma característica típica da categoria; no entanto, o pinguim é um pássaro que não voa. Consequentemente, eliminar a característica voar porque não gera variação significativa dentro da categoria significaria eliminar uma característica distintiva do pinguim (o pinguim é uma ave que não voa).

Sob esses pressupostos, foi proposta uma transformação simples $f(x)$ que comprime a representação semântica do objeto proposta (Ver Figura 6.1e) em uma assinatura semântica de baixa dimensionalidade (Ver Figura 6.1g). A função de transformação proposta visa reduzir a dimensionalidade da representação semântica do objeto, mas mantendo a propriedade –presente na representação inicial– de recuperar facilmente o *significado semântico do objeto* e a *diferença semântica do objeto* a partir da assinatura final do descritor. A assinatura final que representa a descrição semântica global do objeto (ψ_o) é construída concatenando as assinaturas resultantes da transfor-

Algoritmo 3 Redução de Dimensionalidade $f(x)$

1: ***Entrada***: *vetor m-dimensional* α, Ω_i, b_i, *tipo*
2: ***Saída***: *assinatura semântica*
3: ***Parâmetro***: matriz auxiliar $\chi_{r\times r}$
4: $\bar{b_i} \leftarrow \frac{b_i}{m}$ // vetor m-dimensional $\bar{b_i}$ ($b_i = \sum_m \bar{b_i}$)
5: $\chi_{r\times r} \leftarrow shape(r, r)$ // configurar a dimensão da matriz (χ)
6: Encontrar a melhor configuração p, q onde $p \equiv 0 \pmod r$, $q \equiv 0 \pmod r$ e $m = p \cdot q$
7: $\alpha, \Omega_i, \bar{b_i} = reshape\_to\_matrix_{p\times q}(\alpha, \Omega_i, \bar{b_i})$
8: Calcular a matriz de ângulos: $\Theta_{r\times r} = angles\_from(\chi_{r\times r})$
9: $assinatura \leftarrow []$
10: **for** $j = 1, ..., \frac{p}{r}; k = 1, ..., \frac{q}{r}$ **do**
11: Mapear $\chi^{jk}_{r\times r}$ em $\alpha, \Omega_i, \bar{b_i}$
12: Calcular $\vec{z_i}^{jk}$ usando o *produto Hadamard* $\odot$.
13: $\vec{z_i}^{jk} = \begin{cases} \Omega_i^{jk} \odot \alpha^{jk} + \bar{b_i}^{jk}, & \text{se } tipo = significado \\ \left|\Omega_i^{jk}\right| \odot \alpha^{jk}, & em\ outro\ caso \end{cases}$
14: $g^{jk} \leftarrow vetores(\Theta_{r\times r}, \left|\vec{z_i}^{jk}\right|, sign(\vec{z_i}^{jk}))$.
15: $assinatura^{jk}(l) = \sum g^{jk}(\theta), \forall \theta \in \Theta_{r\times r} : \theta_l - 45 < \theta \leq \theta_l$ com $\theta_l = l \cdot \frac{\pi}{4}, \forall l = 1, ..., 8$
16: $assinatura \leftarrow assinatura \oplus assinatura^{jk}$
17: **return** *assinatura*

mação $f(x)$ no *vetor do significado semântico* ($\vec{z}$) e no *vetor diferença semântica* ($\vec{\delta}$), respectivamente (Ver linhas 6-8 do Algoritmo 2).

A Figura 6.2 mostra o fluxo dos principais passos que compõem a função de redução de dimensionalidade $f(x)$ proposta. A transformação proposta usa uma matriz quadrada auxiliar ($\chi_{r\times r}$) como parâmetro para controlar a dimensionalidade da assinatura final do descritor GSDP. Dado os vetores que constituem os vetores semânticos *significado semântico* e *diferença semântica* como entrada na função $f(x)$, o fluxo de passos para construir a assinatura do descritor podem ser resumidos como: **1)** redimensionar os vetores de entrada para a melhor configuração bidimensional de matrizes cujas dimensões ($p \times q$) são múltiplos de r (a dimensão da matriz auxiliar $\chi_{r\times r}$); **2-3)** Calcular a matriz de ângulos ($\Theta_{r\times r}$) a partir dos ângulos formados pela posição de cada célula da matriz auxiliar $\chi_{r\times r}$ com relação ao centro da matriz auxiliar $\chi_{r\times r}$; para garantir a unicidade dos ângulos, os ângulos diagonais foram distribuídos uniformemente entre as magnitudes dos ângulos ς e β; **4-5)** Construir o gradiente semântico unitário para cada matriz mapeada nas matrizes $p \times q$ usando a matriz de ângulos $\Theta_{r\times r}$ como janela deslizante (sem interseções), cada gradiente semântico é construído usando matriz de ângulos $\Theta_{r\times r}$, e a *magnitude* e *sinal* dos vetores semânticos calculados usando as Definições 5 e 9; **6)** Reduzir cada gradiente semântico para oito (8) vetores de forma

semelhante à abordagem SIFT (Lowe, 2004); **7)** Concatenar, para cada gradiente semântico, a correspondente 8D-assinatura unitária resultante do fluxo de passos 4-6. O Algoritmo 3 resume todos os passos da transformação proposta (Ver em Apêndice F mais detalhes de cada passo).

Observa-se que a assinatura semântica final do descritor GSDP (Figura 6.1g) preserva o *significado semântico* do objeto (Ver Propriedade 1) e a *diferença semântica* do objeto (Ver Propriedade 2) presentes na representação semântica inicial das características do objeto (Figura 6.1e). Além disso, dependendo do vetor de entrada na transformação *f(x)* (Ver linhas 6-8 do Algoritmo 2), o descritor GSDP pode usar a transformação *f(x)* para construir representações semânticas (assinaturas) com significados diferentes dentro da categoria do objeto. Por exemplo, a assinatura correspondente ao *protótipo abstrato* da categoria pode ser gerada assumindo como vetor do significado semântico o *vetor semântico do protótipo abstrato* ($M_i \in P_i$) e como vetor *diferença semântica* o vetor zero, pois o protótipo abstrato da categoria não possui diferenças com ele mesmo ($|M_i - M_i| = \vec{0}$). Analogamente, é possível construir uma assinatura para o *protótipo semântico* da categoria (ψ_i), mas usando como *vetor diferença semântica* o vetor resumo de todas as *diferenças semânticas* dos membros representativos da categoria ($\Sigma_i \in P_i$). Ou seja, o descritor GSDP proposto pode construir assinaturas semânticas para: *i)* um objeto, *ii)* o membro ideal da categoria (o protótipo abstrato) e *iii)* o protótipo semântico (Ver Propriedade 3).

6.3 Propriedades do descritor GSDP

Propriedade 1. *Preservação do significado semântico.* A assinatura do descritor GSDP preserva o *significado semântico* (valor semântico) do objeto de entrada:

$$\sum_{k=0}^{|\psi|/2} \psi[k] = z. \tag{6.1}$$

Demonstração. Para demonstrar isso por *prova direta*, basta usar o vetor significado semântico (*tipo=significado*) e seguir em sentido inverso os passos 6 e [10, 17] dos Algoritmos 2 e 3, respectivamente.$\sum_{k=0}^{|\psi|/2} \psi = \sum f\left(A, \Omega_i, b_i, \textit{significado}\right) = \sum \sum_j \sum_k g^{jk} = \sum \sum_j \sum_k \vec{z}^{jk} = \sum \Omega_i \odot \alpha + \bar{b_i} = \sum \vec{z} = z$, com $A \in \{M_i, F_o\}$. □

Propriedade 2. *Preservação da distância prototípica.* A assinatura do descritor GSDP

preserva a *distância prototípica* do objeto de entrada:

$$\sum_{k=|\psi_o|/2}^{|\psi_o|} \psi_o[k] = \delta(o, P_i). \tag{6.2}$$

Demonstração. De maneira semelhante à demonstração anterior, mas usando o vetor diferença semântica do objeto (*tipo=diferença*) e seguindo o sentido inverso dos passos 7 e [10, 17] dos Algoritmos 2 e 3, respectivamente. $\sum_{k=|\psi|/2}^{|\psi|} \psi_o = \sum f\left(|F_o - M_i|, \Omega_i, b_i, \textit{diferença}\right) = \sum \sum_j \sum_k g^{jk} = \sum \sum_j \sum_k \vec{z}^{jk} = \sum |\Omega_i| \odot \alpha = \sum |\Omega_i| \odot |F_o - M_i| = \delta(o, P_i)$. □

Propriedade 3. *Polimorfismo estrutural.* O descritor GSDP proposto possui a propriedade polimórfica de descrever, com a mesma representação estrutural, significados semânticos marcadamente diferentes dentro da categoria. O descritor GSDP possui a habilidade de construir taxonomias de assinaturas semânticas diferentes para:

i) um objeto específico da i-ésima categoria, $\psi_o = \psi(o \in O_{c_i}) = f\left(F_o, \Omega_i, b_i, \textit{significado}\right) \oplus f\left(|F_o - M_i|, \Omega_i, b_i, \textit{diferença}\right) = \psi\left(F_o, |F_o - M_i|, \Omega_i, b_i\right)$;

ii) o protótipo abstrato (*centro semântico abstrato*) da i-ésima categoria, $\psi_{P_i} = f\left(M_i, \Omega_i, b_i, \textit{significado}\right) \oplus f\left(|M_i - M_i|, \Omega_i, b_i, \textit{diferença}\right) = \psi\left(M_i, \vec{0}, \Omega_i, b_i\right)$;

iii) o protótipo semântico (*significado semântico central*) da i-ésima categoria: $\psi_i = f\left(M_i, \Omega_i, b_i, \textit{significado}\right) \oplus f\left(\Sigma_i, \Omega_i, b_i, \textit{diferença}\right) = \psi\left(M_i, \Sigma_i, \Omega_i, b_i\right)$.

6.4 Experimentos e Resultados

Nesta seção são apresentados alguns exemplos dos resultados experimentais realizados para a análise e avaliação das representações construídas com o descritor GSDP proposto. Analisam-se alguns detalhes da construção das assinaturas, além de mostrar exemplos das taxonomias de assinaturas construídas pelo descritor proposto. Apresentam-se também, uma análise da informação semântica contida nas assinaturas projetadas com o intuito de mostrar que preservam o significado semântico contido na *representação semântica do objeto* (Ver Figura 6.1 e)). Na análise semântica realizada mostrou-se que as assinaturas, pela a sua natureza de construção, podem ser interpretáveis e permitem recuperar o significado semântico do objeto necessário para reproduzir as análises semânticas realizadas no Capítulo 5.

6.4.1 Configuração Experimental

Modelos e Banco de protótipos semânticos

Observa-se que a metodologia de descrição semântica proposta (Vide Figura 6.1) possui como requerimentos: *i)* o uso de um modelo CNN base para a extração de características e para a classificação do objeto; e *ii)* o processamento a priori do banco de protótipos semânticos correspondente ao modelo CNN selecionado. Nos experimentos realizados foram usados os mesmos modelos CNN de classificação e banco de imagens usados no Capítulo 5 para calcular os protótipos semânticos correspondentes. Nos experimentos realizados foram usados os modelos simples-MNIST e simples-CIFAR como modelos pilotos para as análises semânticas das representações construídas com o descritor GSDP proposto.

Foram selecionados os modelos CNN de classificação pré-treinados VGG16 (Simonyan & Zisserman, 2014) e ResNet50 (He et al., 2016) como os modelos CNN bases do modelo de descrição semântica global de objetos proposto. O critério de escolha desses modelos justifica-se porque as características extraídas com os modelos VGG16 e ResNet50 são usadas como características primárias em uma variedade de tarefas de processamento semântico da imagem, como detecção de objetos (Ren et al., 2015), anotação de imagens (Murthy et al., 2015), reconhecimento de emoções em vídeo (Xu et al., 2016), transferência de estilo (Gatys et al., 2015), alinhamento de imagens (Han et al., 2017; Rocco et al., 2017, 2018), agrupamento e classificação de cenas (Lu et al., 2017). Observa-se que as características da arquitetura do modelo de descrição semântica baseado em protótipos proposto permite que seja escalável, e possa ser facilmente adaptado a qualquer outro modelo-CNN de classificação que seja escolhido como modelo base.

6.4.2 Interpretação semântica das assinaturas

A semântica na *semântica linguística* constitui o estudo do significado atribuível (ou *interpretável*) de expressões bem formadas. No contexto das matemáticas e a lógica, a semântica refere-se a linguagens formais ou expressões formais cujo significado é *interpretável* em conjuntos de símbolos que cumprem determinadas propriedades abstratas definidas em expressões formais. Nas ciências cognitivas a semântica está relacionada à combinação de signos e à maneira pela qual a mente atribui (ou interpreta) relações permanentes entre essas combinações de signos e outros fatos que não estão naturalmente relacionados a esses símbolos. De forma geral, a semântica está diretamente relacionada com a *interpretação* ou a *atribuição de significados* a vários conceitos,

símbolos, representações ou objetos de interesse.

A motivação principal desta pesquisa reside na ideia de construir representações semânticas que permitam interpretar o significado da informação contida na imagem. Diferente dos trabalhos existentes na literatura para a codificação semântica das imagens de objetos, a abordagem proposta constrói uma representação semântica que permite, por si só, interpretar a informação semântica do objeto que se está descrevendo.

Os experimentos realizados no domínio ds características-CNN mostraram que o *protótipo semântico* da categoria, o *valor semântico* do objeto, e a *distância prototípica* organizam prototipicamente todos os membros da categoria em uma posição específica (e única) dentro da estrutura semântica interna da categoria (Ver Capítulo 5). A ideia por trás do descritor GSDP proposto é encapsular, em uma representação vetorial, a mesma interpretação semântica das características do objeto capturada pelo modelo CPM proposto.

O descritor GSDP proposto, constrói uma *representação semântica do objeto* regida pela informação semântica contida nos protótipos semânticos das categorias (Ver Figura 6.1 e)). Seguidamente, o descritor reduz a grande dimensionalidade da *representação semântica do objeto* em uma assinatura semântica que preserva o *significado semântico* do objeto (Ver Propriedade 1) e a *distância prototípica* do objeto (Ver Propriedade 2). As Propriedades 1 e 2 do descritor proposto permitem recuperar facilmente a informação semântica do objeto (*valor semântico* e *distância prototípica*). Também, o descritor GSDP consegue construir uma assinatura semântica para cada *protótipo abstrato* das categorias de objetos (Ver Propriedade 3).

Doravante apresenta-se como o descritor GSDP consegue codificar e preservar as informações semânticas contidas nas características CNN do objeto (valor semântico e distância prototípica). Mostra-se também como recuperar, das assinaturas do descritor GSDP, essas informações semânticas para reconstruir a estrutura semântica interna da categoria (organização prototípica) alcançada no domínio das características CNN (Ver seção 5.6.4) .

Organização Prototípica no domínio $\mathbb{R}^2$

Similar à abordagem usada no espaço métrico das características-CNN dos objetos (Ver Capítulo 5), visualizar a estrutura interna de uma categoria específica, no espaço das características-GSDP, consiste na visualização do conteúdo semântico das assinaturas GSDP de cada elemento da categoria. Observa-se que no espaço das características GSDP dos elementos da i-ésima categoria (ψ_{c_i}), a função de distância L1 constitui uma métrica; e consequentemente (ψ_{c_i}, L_1) é um espaço métrico.

Proposição 2. Sejam os *espaços métricos* (ψ_{c_i}, L_1) e $(\mathbb{R}^2, L_1)$ onde ψ_{c_i} representa o domínio das assinaturas GSDP dos objetos que pertencem à i-ésima categoria $c_i \in C, \forall i = 1...n$. A função $\lambda : (\psi_{c_i}, L_1) \to (\mathbb{R}^2, L_1) \mid \lambda(\psi_o \in \psi_{c_i}) = p(\sum_{k=0}^{|\psi|/2} \psi_o[k], \sum_{|\psi|/2}^{|\psi|} \psi_o[k]) = p(z_o, \delta(o, P_i))$ é uma função contínua.

Demonstração. De maneira similar à demonstração da Proposição 1 é simples demonstrar, usando as Propriedades 1 e 2 das assinaturas-GSDP, que a função $\lambda : (\psi_{c_i}, L_1) \to (\mathbb{R}^2, L_1)$ é uma função contínua (Ver exemplo das relações de distâncias entre esses espaços métricos no Apêndice G.2). □

Assim, similar à abordagem usada para visualizar a categoria no espaço m-dimensional das características-CNN, é suficiente visualizar no espaço métrico $(\mathbb{R}^2, L_1)$ a informação contida nas assinaturas-GSDP dos membros da i-ésima categoria mapeadas através da função contínua $\lambda : (\psi_{c_i}, L_1) \to (\mathbb{R}^2, L_1)$. Consequentemente, o comportamento observado –em termos de distância– da i-ésima categoria (c_i) mapeada no espaço métrico $(\mathbb{R}^2, L_1)$, é equivalente ao comportamento observado na estrutura interna da categoria no espaço métrico das assinaturas GSDP (ψ_{c_i}, L_1).

A Figura 6.3 mostra o comportamento da estrutura semântica interna da categoria *número cinco* (c_5) do banco de dados MNIST. Apresentam-se as representações

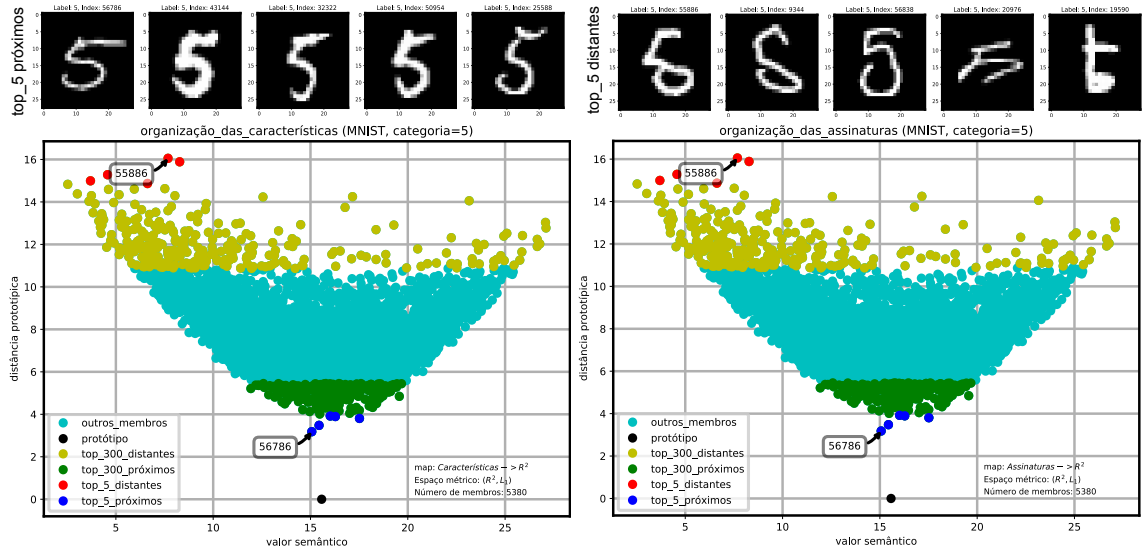


Figura 6.3: *Organização prototípica dentro da categoria c_5 do banco de dados MNIST.* No topo, da esquerda para a direita, os 5 elementos mais próximos e os 5 elementos mais distantes do protótipo semântico P_5 na categoria c_5. O índice do primeiro elemento é anotado dentro da caixa preta. Observa-se como o domínio das assinaturas preserva a disposição interna da categoria alcançada pelas características no domínio $\mathbb{R}^2$.

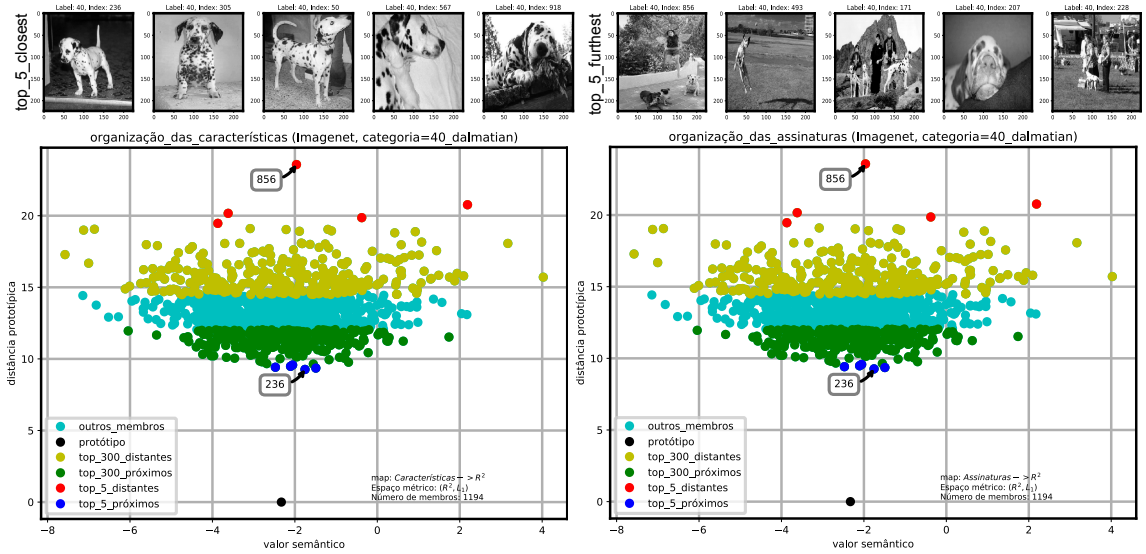


Figura 6.4: *Organização prototípica dentro da categoria c_{40}(dalmatian) no banco de dados ImageNet.* No topo, da esquerda para a direita, os 5 elementos mais próximos e os 5 elementos mais distantes do protótipo semântico P_{40} na categoria c_{40}.

alcançadas pela categoria mapeada usando como origem o domínio das características-CNN do objeto (F_{c_5})(Esquerda) (Ver Capítulo 5) e usando o domínio das assinaturas-GSDP correspondentes (ψ_{c_5})(Direita). Foi comparado o comportamento da estrutura semântica interna da categoria em ambos os domínios. Observa-se como a função contínua $\lambda : (\psi_{c_i}, L_1) \rightarrow (\mathbb{R}^2, L_1)$ consegue mapear a informação semântica contida nas assinaturas-GSDP dos membros da categoria para, exatamente, a posição semântica atingida pelas características-CNN do objeto no domínio $\mathbb{R}^2$.

O anterior significa que os 5 elementos mais próximos (pontos em azul) e os 5 elementos mais distantes (pontos em vermelho) da assinatura do protótipo semântico (ψ_{P_5}) são mapeados na mesma posição semântica (no domínio $\mathbb{R}^2$) que as características dos objetos correspondentes. Ou seja, as características extraídas dos objetos e as assinaturas do descritor correspondentes possuem a mesma posição semântica no domínio $\mathbb{R}^2$. De maneira similar, a visualização dos top-300 elementos mais próximos e dos top-300 elementos mais distantes em diferentes cores permite perceber a mesma organização prototípica na estrutura interna da categoria c_5. As assinaturas dos objetos típicos são posicionadas semanticamente próximas da assinatura do protótipo (ψ_{P_5}), e as assinaturas dos objetos menos representativos são encontradas distantes do centro semântico da categoria. O Apêndice H apresenta outros exemplos da organização prototípica nas categorias dos bancos de dados analisados.

As Figuras 6.4 e 6.5 mostram exemplos da organização prototípica alcançada pela abordagem proposta na estrutura interna das categorias c_{40}-*dalmatian* e c_9-*Persian*

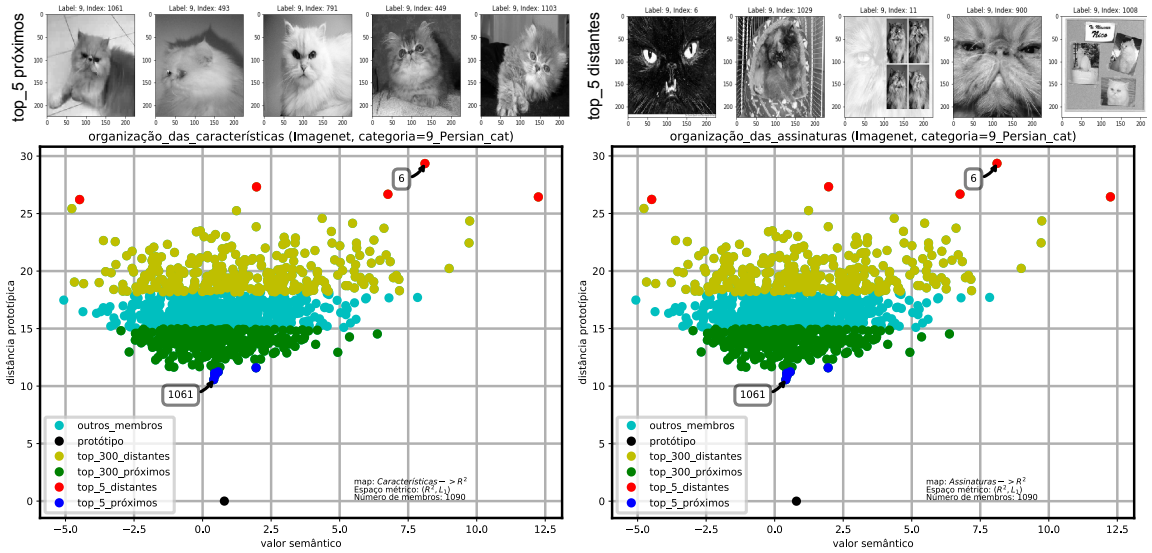


Figura 6.5: *Organização prototípica dentro da categoria c_9(Persian cat) no banco de dados ImageNet.* Mostram-se os 5 elementos mais próximos e os 5 elementos mais distantes do protótipo semântico P_9 na categoria c_9.

cat do banco de dados ImageNet usando o modelo VGG16. Observa-se que mapeando o conteúdo das assinaturas-GSDP dos objetos da i-ésima categoria (ψ_{c_i}), além de formar uma organização prototípica dos elementos dentro da categoria no domínio $\mathbb{R}^2$, preserva-se a mesma estrutura semântica interna da categoria gerada pelas características-CNN mapeadas. Nos exemplos apresentados é detalhado o tipo de mapeamento realizado e a quantidade de elementos mapeados da categoria.

No Capítulo 5 demonstrou-se que o *valor semântico* e a *distância prototípica* definem a posição semântica que distingue às características do objeto dentro da categoria. Os resultados apresentados nesta seção mostram que o *valor semântico* do objeto e a *distância prototípica* encontram-se encapsulados na assinatura do descritor GSDP proposto. A codificação semântica de imagens de objetos proposta distingue-se das abordagens da literatura, porque as assinaturas do descritor semântico proposto encapsulam, usando o *protótipo semântico* da categoria, o significado das características do objeto usado pelo modelo CPM para representar o objeto dentro da categoria. As assinaturas GSDP possuem a propriedade de ser interpretáveis e permitem recuperar o mesmo significado semântico das características do objeto (*valor semântico*) usado pelos modelos CNN para a classificação do objeto. A representação semântica proposta permite facilmente recuperar a *distância prototípica* do objeto ($\delta(o, P_i)$) e consequentemente ter, baseado na assinatura-GSDP do objeto, uma pontuação da tipicidade do objeto dentro da categoria (*pontuação de tipicidade* $(o) = 1/\delta(o, P_i)$).

6.4.3 Assinaturas do descritor

Comprimento das assinaturas

Cada modelo CNN de classificação possui uma dimensionalidade específica ($|F|$) na camada totalmente conectada usada como característica do objeto. Consequentemente, os vetores que compõem os protótipos semânticos calculados possuem a mesma dimensionalidade que as características extraídas com o modelo CNN usado como modelo-base. Assim, o comprimento das assinaturas do descritor GSDP depende das dimensionalidades da característica extraída e da matriz unitária $\chi_{r\times r}$ usada como parâmetro na transformação de redução de dimensionalidade proposta $f(x)$ para comprimir a representação semântica do objeto (Ver Figura 6.2 e Algoritmo 3).

A Tabela 6.1 mostra o comprimento das assinaturas do descritor GSDP proposto para cada um dos modelos CNN de classificação usados como modelo-base. Para cada modelo CNN usado, foram projetados – mudando a dimensionalidade da matriz auxiliar $\chi_{r\times r}$ – dois comprimentos de assinaturas do descritor GSDP diferentes. Como foi apresentado no fluxo de passos da transformação $f(x)$ (Ver Figura 6.2), cada vetor de característica F de entrada é transformado para uma nova geometria $F_{p\times q}$ onde é computado cada gradiente semântico (g^{jk}) usando como janela deslizante a matriz de ângulos ($\Theta_{r\times r}$) construída a partir da matriz auxiliar $\chi_{r\times r}$. A partir da matriz de gradientes resultantes ($M_{g^{jk}}$), a assinatura semântica resultante de $f(x)$ é construída através da concatenação dos 8D-vetores unitários calculados usando cada gradiente semântico que compõe a matriz de gradientes. O comprimento da assinatura final do descritor ($|\psi|$) possui duas vezes a dimensionalidade das representações intermédias construídas com a transformação $f(x)$ (Ver Algoritmo 2 linha 8). A Tabela 6.1 apresenta os detalhes das configurações usadas para construir as assinaturas do descritor GSDP proposto.

Taxonomias das assinaturas

A Figura 6.6 mostra um exemplo das diferentes taxonomias de assinaturas construídas com o descritor semântico GSDP usando o modelo simples-MNIST como modelo-base (nesse caso as assinaturas GSDP têm tamanho 32). Mostra-se, visualmente, a Propriedade 3 do descritor GSDP proposto de construir assinaturas: do *centro semântico abstrato* da categoria (assinatura do protótipo abstrato) (ψ_{P_5}), do *significado semântico central da categoria* (assinatura do protótipo semântico) (ψ_5) e do significado semântico dos objetos (assinatura de um objeto específico) (ψ_o).

A *assinatura da categoria* pode ser entendida como a assinatura do descri-

Modelo - CNN	$\chi_{r\times r}$	$\lvert F\rvert$	$F_{p\times q}$	g^{jk}	$\lvert f(x)\rvert$	Assinatura $\lvert\psi\rvert$
simples-MNIST	$\chi_{8\times8}$	128	$F_{16\times8}$	2×1	16	32
simples-MNIST	$\chi_{4\times4}$	128	$F_{16\times8}$	4×2	64	128
simples-CIFAR	$\chi_{8\times8}$	512	$F_{32\times16}$	4×2	64	128
simples-CIFAR	$\chi_{4\times4}$	512	$F_{32\times16}$	8×4	256	512
ResNet50	$\chi_{16\times16}$	2048	$F_{64\times32}$	4×2	64	128
ResNet50	$\chi_{8\times8}$	2048	$F_{64\times32}$	8×4	256	512
VGG16	$\chi_{16\times16}$	4096	$F_{64\times64}$	4×4	128	256
VGG16	$\chi_{8\times8}$	4096	$F_{64\times64}$	8×8	512	1024

Tabela 6.1: *Dimensões das assinaturas do descritor GSDP proposto para cada modelo CNN base usado.* Apresentam-se detalhes das configurações usadas para construir as assinaturas do descritor GSDP proposto. Mostram-se: as dimensões da matriz auxiliar ($\chi_{r\times r}$); a dimensionalidade das características ($|F|$); a nova geometria da característica ($F_{p\times q}$); a quantidade de gradientes unitários (g^{jk}); o comprimento ($|f(x)|$) das assinaturas construídas com a transformação $f(x)$; e a dimensão final ($|\psi|$) das assinaturas construídas com o descritor GSDP proposto. Fonte: Elaborado pelo autor.

tor semântico de uma categoria específica. A *assinatura da categoria* é construída usando o Algoritmo 2 a partir da informação semântica contida (e resumida) no *protótipo semântico* da categoria. Por exemplo, a assinatura da c_5-categoria (ψ_5) mostrada na Figura 6.6 é construída usando a informação contida no protótipo da c_5-categoria ($P_5 = (M_5, \Sigma_5, \Omega_5, b_5)$). A assinatura da c_5-categoria ($\psi_5 = \psi\left(M_5, \Sigma_5, \Omega_5, b_5\right) = f\left(M_5, \Omega_5, b_5, \textit{significado}\right) \oplus f\left(\Sigma_5, \Omega_5, b_5, \textit{diferença}\right)$) é construída pela concatenação da assinatura gerada a partir do *protótipo abstrato* da c_5-categoria (M_5) e pela assinatura gerada a partir do vetor constituído pelos valores desvio padrão de todas as *diferenças semânticas* dos objetos da categoria (Σ_5) (Ver Algoritmo 2 e Propriedade 3). A segunda metade da *assinatura da categoria* pode ser entendida como a distribuição de valores (ou *valores fronteira do protótipo*) que regem os limites máximos das assinaturas construídas a partir do *vetor diferença semântica* dos objetos. O anterior significa que os elementos mais típicos da categoria terão geralmente, na assinatura correspondente, uma distribuição dos valores da diferença semântica menor que a distribuição dos valores que regem a fronteira do protótipo. No exemplo apresentado na Figura 6.6, pode-se observar como o elemento mais típico da categoria c_5 (id 56786) possui na assinatura uma distribuição que cumpre com o expressado anteriormente.

A *assinatura do protótipo abstrato* encapsula a informação que representa o centro semântico abstrato da categoria e constitui uma versão degenerada da *assinatura da categoria*. Por exemplo, na Figura 6.6 a assinatura do protótipo abstrato da c_5-

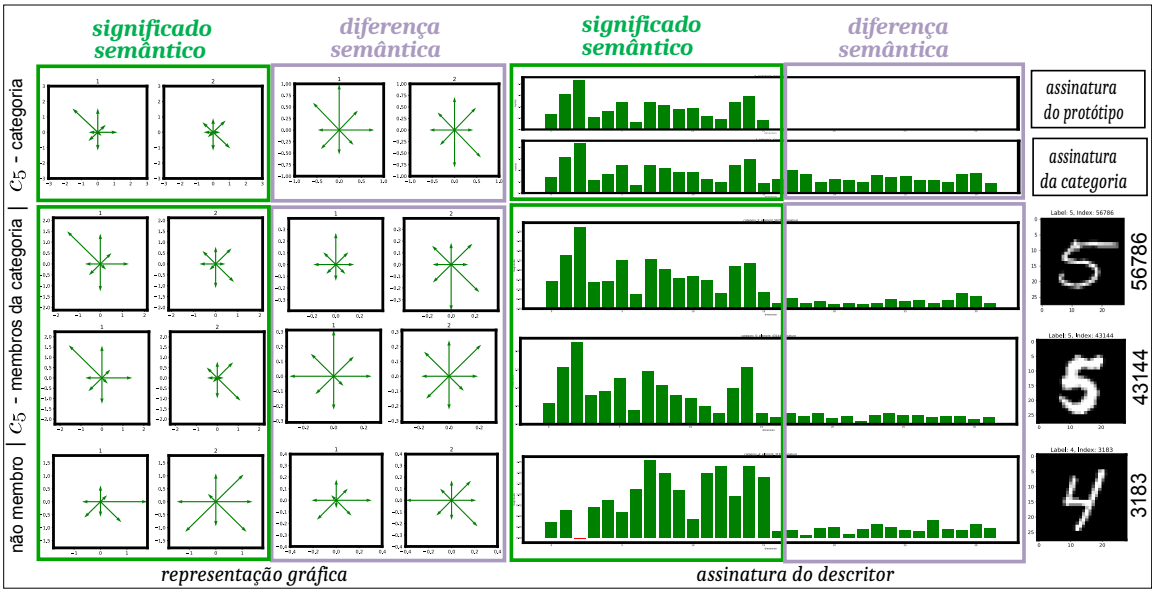


Figura 6.6: *Taxonomias das assinaturas semânticas construídas com o descritor GSDP proposto para a c_5-categoria do banco de dados MNIST.* Mostram-se exemplos - da categoria c_5 do banco de dados MNIST - das possíveis assinaturas semânticas que podem ser construídas com o descritor proposto: a assinatura do *protótipo abstrato da categoria*, a assinatura do *protótipo semântico da categoria* (assinatura da categoria) e a assinatura de um objeto. Apresentam-se exemplos de assinaturas de dois membros da c_5-categoria e de um membro que não pertence à c_5-categoria.

categoria (ψ_{P_5}) pode ser representada como a versão degenerada da assinatura da c_5-categoria ($\psi_{P_5} = \psi(M_5, \vec{0}, \Omega_5, b_5) = f(M_5, \vec{0}, b_5, \textit{significado}) \oplus \vec{0}$). O *protótipo abstrato* da categoria pode ser entendido como o elemento ideal da categoria ou a distribuição de valores das características (ou sequência DNA) que distingue aos elementos que pertencem à categoria. Consequentemente, a *assinatura do protótipo abstrato* pode ser entendida como a distribuição de valores (ou assinatura-DNA da categoria) que representa à família de assinaturas dos objetos que pertencem à categoria. Os membros da categoria terão na assinatura do descritor GSDP correspondente, uma representação semântica similar à assinatura-DNA da categoria.

A *assinatura de um objeto* da categoria (ψ_o) é construída usando o Algoritmo 2 a partir das características do objeto (F_o), as *diferenças semânticas* do objeto com relação ao protótipo da categoria ($|F_o - M_i|$) e a relevância das características na categoria (Ω_i). A assinatura do objeto (ψ_o) encapsula na primeira metade o *significado semântico* do objeto (Ver Propriedade 1) e na segunda metade a *distância prototípica* do objeto (Ver Propriedade 2).

Por exemplo, na Figura 6.6 observa-se como a assinatura de um elemento (id

56786) representativo da c_5-categoria ($\psi_{o56786} = \psi\left(F_{o56786}, |F_{o56786} - M_5|, \Omega_5, b_5\right) = f\left(F_{o56786}, \Omega_5, b_5, \textit{significado}\right) \oplus f\left(|F_{o56786} - M_5|, \Omega_5, b_5, \textit{diferença}\right)$) é similar à assinatura do protótipo abstrato da categoria. Note-se que os membros da c_5-categoria apresentados na Figura 6.6 constituem elementos próximos ao protótipo semântico da c_5-categoria (pertencem ao Top-5 mais próximos). Consequentemente, o *significado semântico* representado nas assinaturas de objetos relevantes da categoria será muito semelhante ao *significado semântico* representado na assinatura do protótipo abstrato da categoria. Aliás, a representação da *diferença semântica*, nas assinaturas desses membros da categoria, será mais próxima ao vetor zero ($\vec{0}$) enquanto maior seja a tipicidade do objeto (menor *distância prototípica*). A abordagem proposta permite que membros representativos da categoria possuam assinaturas semelhantes à assinatura do protótipo abstrato da categoria, e que membros que não pertencem à categoria possuam assinaturas marcadamente diferentes (Exemplo: Na Figura 6.6 membro id=3183).

Representação gráfica das assinaturas

Para uma melhor análise das assinaturas do descritor GSDP proposto, as assinaturas do descritor podem ser visualizadas em versão gráfica ou versão histograma (vetor dos valores da assinatura). A representação gráfica das assinaturas do descritor GSDP proposto está constituída pelos eixos com a orientação, sentido, e magnitude dos 8D-vetores resultantes de cada gradiente semântico construído.

A Figura 6.6 mostra exemplos das representações (versão gráfica e assinatura correspondente) do descritor GSDP para o modelo simples-MNIST no banco de dados MNIST. Observa-se que as características extraídas com o modelo simples-MNIST possuem dimensionalidade 128 e, consequentemente, as assinaturas GSDP_MNIST possuem dimensionalidade 32 para o modelo simples-MNIST (versão que usa a matriz auxiliar $\chi_{8\times8}$) (Ver Tabela 6.1). Em cada execução da transformação $f(x)$ (*significado semântico* e *diferença semântica* do objeto) são construídas assinaturas intermédias que possuem uma 16D-dimensionalidade (2 gradientes semânticos). Consequentemente, as assinaturas do descritor possuem dimensionalidade 32 e a representação gráfica da assinatura está composta pelos quatro 8D-vetores.

A Figura 6.7 mostra um exemplo das diferentes taxonomias de assinaturas construídas com o descritor GSDP proposto usando o modelo VGG16 no banco de dados ImageNet (assinaturas GSDP com tamanho 256 nesse caso). Nota-se que as características extraídas com o modelo VGG16 possuem dimensionalidade 4096, e consequentemente as assinaturas GSDP_VGG16 possuem dimensionalidade 256 (versão que usa a matriz auxiliar $\chi_{16\times16}$) (Ver Tabela 6.1). Nesse caso, em cada execução da transfor-

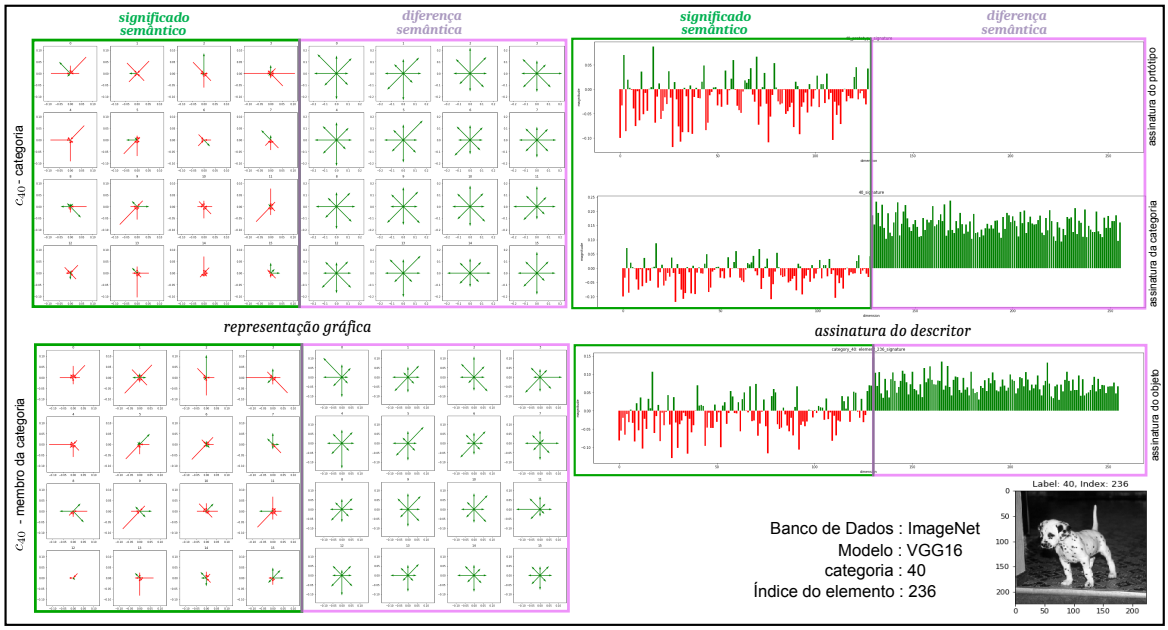


Figura 6.7: *Exemplos das assinaturas do Descritor Semântico Global proposto na categoria c_{40}(n02110341-dalmatian) do banco de dados ImageNet usando o modelo VGG16.* Mostra-se a assinatura do protótipo, a assinatura do significado semântico da categoria c_{40} e a assinatura do objeto com índice 236 na categoria.

mação $f(x)$ (*significado semântico* e *diferença semântica* do objeto) são construídas assinaturas intermédias que possuem uma 128D-dimensionalidade (16 gradientes semânticos). Consequentemente, as assinaturas do descritor possuem dimensionalidade 256 e a representação gráfica da assinatura está composta pelos 16 8D-vetores construídos dos gradientes semânticos.

Estrutura interna

O descritor GSDP proposto tenta construir - usando o protótipo semântico da categoria - uma família (ou distribuição) de assinaturas específica para cada categoria de objetos. O intuito dessa abordagem é conseguir que a estrutura interna das assinaturas-GSDP dos membros da categoria tenham uma representação semelhante à assinatura-GSDP do protótipo abstrato da categoria; e que membros que não pertencem à categoria possuam representações marcadamente diferentes. Como se apresentou em seções anteriores, o descritor GSDP consegue que elementos muito típicos tenham assinaturas do descritor semelhantes à assinatura do protótipo abstrato; e que elementos que não pertencem à categoria tenham uma assinatura com uma estrutura interna muito diferente das assinaturas-GSDP desses membros representativos da categoria.

Os experimentos também mostraram que os elementos da categoria podem ser

agrupados por sua semelhança semântica dentro da categoria (*family resemblance*). O anterior é consequência de que as assinaturas GSDP encapsulam o significado capturado pelo modelo CPM proposto. Assim, os membros representativos da categoria são agrupados próximos ao protótipo da categoria e outros menos representativos são posicionados mais distantes desse centro semântico.

Mesmo quando esses resultados são relevantes, essas representações da estrutura semântica interna das categorias descrevem –somente– o comportamento da informação contida nas representações dos *membros* da categoria. Ou seja, como na estrutura interna somente são representados membros da categoria de interesse, isso não significa que não possam existir membros de outras categorias com representações semelhantes à estrutura interna das assinaturas-GSDP da categoria. O anterior significa que no espaço m-dimensional das assinaturas-GSDP (e das características-CNN correspondentes) podem existir elementos de outra categoria na vizinhança de membros de uma categoria específica. Usou-se o algoritmo de visualização t-SNE (Maaten & Hinton, 2008) para analisar a vizinhança dos elementos no espaço m-dimensional das representações construídas, porque esse algoritmo consegue preservar a estrutura local entre os elementos de uma categoria. O algoritmo t-SNE expõe que os elementos próximos uns dos outros no conjunto de dados de alta dimensão, tenderão estar próximos uns dos outros no mapa de baixa dimensão da visualização t-SNE.

Analisou-se o poder discriminativo e o desempenho na visualização t-SNE da representação semântica proposta para as imagens de objetos (GSDP) *versus* as características da imagem extraídas usando modelos CNN de classificação. Para cada modelo CNN usado como modelo base pelo descritor GSDP proposto, comparou-se o desempenho t-SNE da família de características construídas com cada modelo CNN usado. A visualização t-SNE foi realizada para a família de características constituídas pelas: características CNN, as assinaturas GSDP correspondentes, e as versões das características CNN reduzidas usando PCA (Abdi & Williams, 2010) (as características CNN foram reduzidas para a mesma dimensionalidade das assinaturas GSDP).

A Figura 6.8 mostra a visualização t-SNE de 10 categorias dos banco de dados de imagens usados, para cada um dos modelos CNN disponíveis no descritor GSDP proposto. Usou-se a distância euclidiana como medida de similaridade entre as representações construídas e 50 como o valor de perplexidade do algoritmo t-SNE (Ver Apêndice I para outros exemplos). Na Figura 6.8 observa-se como as representações-GSDP são agrupadas de forma compacta e com maior separação entre as categorias do que as agrupações construídas com as características dos modelos CNN (e suas versões reduzidas usando PCA). Esse resultado evidencia que o descritor GSDP constrói representações semânticas da imagem do objeto com estruturas internas semelhantes

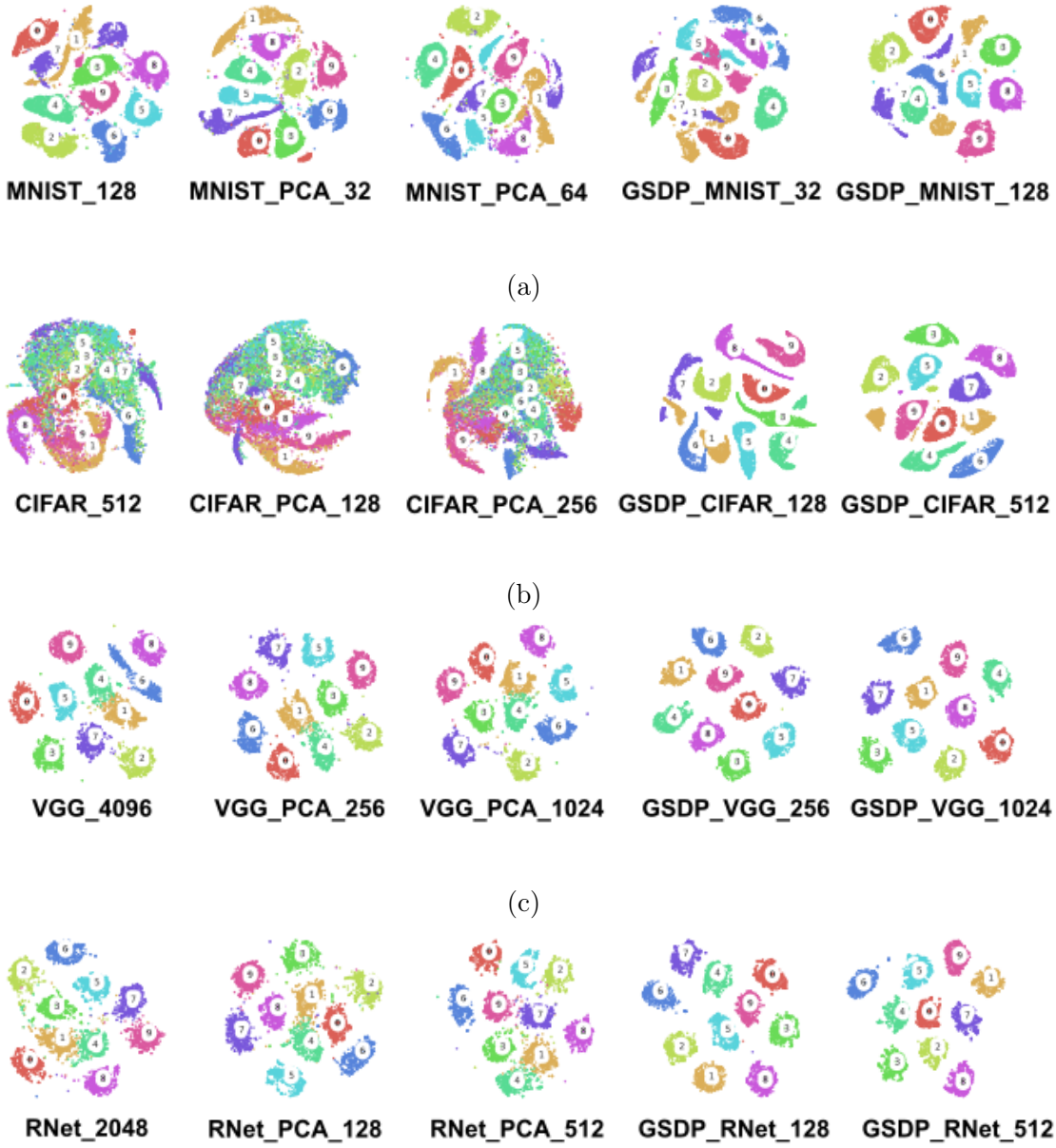


Figura 6.8: *Visualização t-SNE. a)* visualização t-SNE das características construídas com o modelo *simples-MNIST* no conjunto de dados MNIST; *b)* visualização t-SNE das características construídas com o modelo *simples-CIFAR* no conjunto de dados CIFAR10; *c,d)* visualização t-SNE das 10 primeiras categorias do conjunto de dados ImageNet usando características construídas com os modelos *VGG16* e *ResNet50*, respectivamente. O comprimento de cada característica (em cada família de características) mostra-se na legenda correspondente.

para elementos da mesma vizinhança local (e da mesma categoria) e representações discriminativamente diferentes para elementos fora da vizinhança local dos membros da categoria. Portanto, pode-se supor que o descritor GSDP proposto constrói representações semânticas da imagem do objeto com a capacidade de maximizar as diferenças das representações dos elementos entre categorias (*inter-class difference*) e minimizar as diferenças das representações intraclasse (*intra-class difference*).

6.4.4 Avaliação do desempenho

Neste Capítulo apresentou-se a metodologia de descrição semântica global de imagens de objetos baseada em protótipos. Observa-se que a natureza global do descritor GSDP proposto impede que seja utilizado -por si só- em tarefas de correspondência semântica; pelo que não pode ser comparado com os descritores semânticos existentes para essas tarefas (Choy et al., 2016; Zhou et al., 2016; Kim et al., 2017; Rocco et al., 2018). Mas, a representação GSDP proposta pode ser avaliada em outras tarefas onde a descrição global da imagem do objeto possui grande relevância.

Embora a análise realizada na taxonomia interna das assinaturas do descritor mostra que a representação GSDP proposta consegue encapsular a informação semântica do objeto representada com o modelo CPM proposto, é necessário avaliar o desempenho da codificação proposta em tarefas comuns de processamento da imagem. Nesta seção foi realizada uma avaliação do desempenho da codificação semântica GSDP em outras tarefas práticas onde a descrição global das características da imagem é utilizada. Nos experimentos realizados usaram-se as assinaturas GSDP para avaliar seu desempenho em métodos de aprendizagem supervisionado e não supervisionado.

Agrupamento

Várias abordagens de análise e processamento semântico da imagem usam como característica semântica a informação semântica contida na última camada totalmente conectada dos modelos CNN de classificação. As características extraídas com esses modelos CNN de classificação, além de alcançar um excelente desempenho na classificação da imagem em grande escala, também são usadas para melhorar o desempenho de outras tarefas (Ren et al., 2015; Murthy et al., 2015; Xu et al., 2016; Gatys et al., 2015; Han et al., 2017; Rocco et al., 2017; Lu et al., 2017). A codificação semântica GSDP proposta baseia-se nas características extraídas desses modelos CNN de classificação (por exemplo VGG16, ResNet50). Uma análise admissível para avaliar o desempenho da codificação semântica proposta reside na comparação da qualidade

e do poder discriminativo da representação GSDP com relação ás características da camada totalmente conectada dos modelos CNN usados como modelo-base.

Alguns trabalhos como o realizado por Yang et al. (2016) mostraram que a qualidade das representações das características extraídas influencia nas tarefas de agrupamento de imagens por similaridade (*clustering*). Yang et al. (2016) aprenderam representações a partir das características da última camada dos modelos CNN, e mostraram que quando essas representações das características aprendidas conseguem boas métricas nas tarefas de agrupamento, podem-se generalizar bem quando são transferidas para outras tarefas. Aliás, Kaufman & Rousseeuw (2009) mostraram que as escalas de cada uma das características unitárias que compõem a representação das características da imagem também possuem influência na tarefa de agrupamento de imagens.

Sob esses pressupostos, uma análise admissível para avaliar a codificação semântica do descritor GSDP proposto reside em verificar a sua utilidade e adequação em tarefas de agrupamento de imagens. Nesta seção são apresentados alguns experimentos da abordagem utilizada para avaliar o desempenho e o poder discriminativo da representação proposta em termos de métricas de agrupamento.

Na avaliação realizada foi comparado o desempenho, em tarefas de agrupamento de imagens, das versões do descritor *GSDP* baseadas nos modelos VGG16 e ResNet50 com relação ao desempenho de outras representações globais. Assim, foram selecionadas como representações globais *handcraft* da imagem os descritores: GIST (Oliva & Torralba, 2001), LBP (Ojala et al., 2002), HOG (Dalal & Triggs, 2005a), Color64 (Li, 2007), Color_Hist (Song et al., 2004), Hu_Haralick_CH (Haralick et al., 1973; Hu, 1962; Song et al., 2004). Na avaliação foram incluídas também descrições *deep-learning* da imagem como as características VGG16 (Simonyan & Zisserman, 2014), ResNet50 (He et al., 2016) e versões PCA (reduzidas com o algoritmo PCA) dessas representações. A avaliação da qualidade dos resultados de agrupamento realizou-se baseado nas métricas de validação de índices externos (Wang et al., 2009). Essa medida de concordância (índice externo) permite avaliar os resultados do algoritmo de agrupamento com base na estrutura de agrupamento conhecida, a priori, do conjunto de dados (os rótulos das categorias).

Os experimentos foram realizados usando o algoritmo K-Means (MacQueen et al., 1967; Arthur & Vassilvitskii, 2006). O critério de escolha desse algoritmo baseia-se em que, além de ser o mais simples e conhecido dos métodos de agrupamento de imagens que usam a abordagem de *particionamento*, o algoritmo possui algumas semelhanças com a representação semântica do descritor proposto. O método K-Means minimiza a soma dos erros quadrados entre os pontos de dados e seus centros de agrupamentos mais próximos. Essa abordagem possui semelhanças com a representação GSDP proposta,

pois as assinaturas do descritor foram construídas com o propósito de organizar as categorias usando como centro da organização (ou agrupamento) o protótipo abstrato da categoria.

No processo de comparação do desempenho das representações (assinaturas) na tarefa de agrupamento, foram extraídas –usando os descritores globais selecionados– as características das imagens que compõem as primeiras 100 categorias dos banco de dados usados. Os descritores foram agrupados em duas categorias: *i)* descritores artesanais (*handcraft features*), e *ii)* descritores aprendidos com aprendizagem profundo (*deep features*). Os experimentos foram realizados em dois banco de dados diferentes: ImageNet para avaliar o desempenho de nossa proposta nas mesmas condições de treinamento das outras representações usadas; e Coco (Lin et al., 2014) para avaliar o desempenho e capacidade de generalização de cada representação de imagem em dados nunca vistos (*crossdataset*). O conjunto de dados Coco é um desafio para as representações de imagens treinadas no ImageNet porque as categorias entre ambos os dois conjuntos de dados não são mapeadas uma a uma (*one-to-one*).

O algoritmo de agrupamento K-Means foi executado, separadamente, nos conjuntos de características extraídas com os descritores globais escolhidos. Para cada grupo de assinaturas resultantes, o experimento foi conduzido de forma incremental, começando com o agrupamento das características correspondentes a três (3) categorias do banco de dados (500 imagens $\times$ categoria). O algoritmo *K-Means* é executado para agrupar essas características na mesma quantidade de partições que as categorias presentes (3). Em seguida, é acrescentado o conjunto de características inicial com as características de outra categoria. O K-Means é executado novamente para agrupar o novo conjunto de características com uma partição a mais (3+1=4). Esse procedimento é executado tantas vezes como o número de categorias máximo (100) selecionado para realizar o experimento.

As métricas de agrupamento (*Homogeneity, Completeness, V-measure, Adjusted Rand Index, Adjusted Mutual Information*)[1] foram anotadas em cada execução do algoritmo K-Means para cada conjunto de assinaturas dos descritores globais usados. O experimento visa observar o comportamento das métricas de agrupamento na medida que aumenta o número de características a serem agrupadas e o número de partições (*clusters*). O intuito do experimento é avaliar o desempenho K-Means das representações (em termos de métricas de agrupamento) na medida que aumenta o volume de dados e a variedade da composição dos dados (número de categorias). As características que possuam melhor poder discriminativo conseguirão um melhor

[1]Doravante usam-se os termos em inglês para evitar ambiguidades conceituais.

desempenho na tarefa de agrupamento de imagens (Yang et al., 2016).

A homogeneidade (*Homogeneity*) (Rosenberg & Hirschberg, 2007) é uma medida que avalia se cada partição (*cluster*) construída contém apenas membros de uma única categoria. A completude (*Completeness*) (Rosenberg & Hirschberg, 2007) é uma métrica que mede se todos os membros de uma determinada categoria são atribuídos à mesma partição. Aliás, a medida *V-measure* (Rosenberg & Hirschberg, 2007) baseia-se nos valores das métricas anteriores e pode ser definida como a média harmônica entre homogeneidade e completude ($v\_measure = 2 \times \frac{(homogeneity \times completeness)}{(homogeneity + completeness)}$). *Adjusted Rand Index (ARI)* (Hubert & Arabie, 1985) e *Adjusted Mutual Information (AMI)* (Vinh et al., 2010) são métricas mais complexas, mas em termos gerais, estão relacionadas com a *acurácia* e a *variação da informação* –respetivamente– dos agrupamentos realizados. Todas as métricas que foram avaliadas possuem valores entre 0-1, alcançando valor 1 quando as partições realizadas são idênticas às partições anotadas conhecidas a priori.

A Tabela 6.2 mostra uma captura (*screenshot*) dos valores das métricas de agrupamento alcançadas pelas representações dos descritores globais selecionados na iteração 20 (primeiras 22 categorias) dos experimentos realizados. Nesse estado (iteração 20) cada instância do experimento agrupa 11000 imagens (500×22) em 22 partições baseado no conteúdo das representações construídas com cada descritor global avaliado. Para cada representação usada como modelo-CNN-base foi construída uma família de representações constituída: características-CNN, versões PCA de menor dimensionalidade dessas características, e as correspondentes versões GSDP com a mesma dimensionalidade que as características PCA (Exemplo:VGG, VGG_PCA_X, GSDP_VGG_X). Essa configuração do experimento também permite avaliar o desempenho da função de redução de dimensionalidade do descritor GSDP em comparação com os resultados alcançados com a abordagem de redução PCA. A Tabela 6.2 também apresenta, para cada abordagem, o tamanho da característica e a velocidade de extração da característica (frames x segundo - FPS).

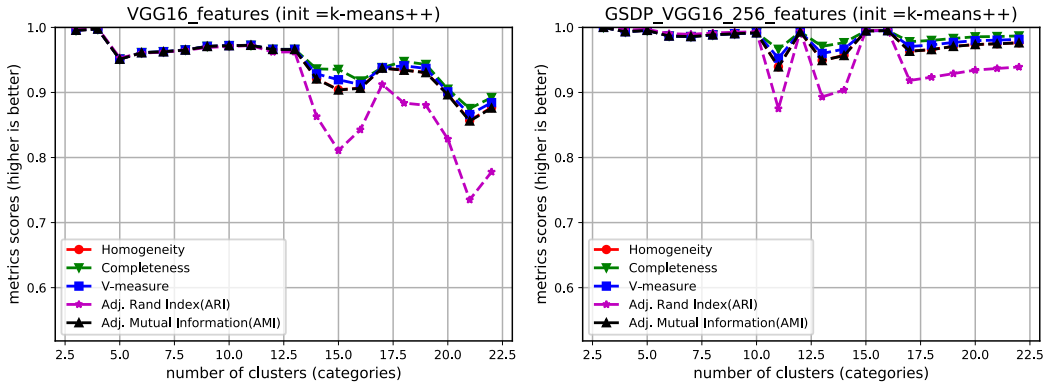
Observa-se nos resultados apresentados na Tabela 6.2 que, nesse estado do experimento, somente as famílias de representações baseadas em modelos-CNN (VGG16, ResNet, GSDP) conseguem boas métricas de agrupamento de imagens. Devido a que esse comportamento é observado durante todo o experimento, doravante foram analisados em detalhe somente aqueles descritores que alcançaram melhor desempenho na tarefa de agrupamento de imagens (Ver o histórico do comportamento das métricas de todos os descritores globais avaliados no Apêndice J).

As Figuras 6.9 e 6.10 mostram exemplos do histórico do comportamento das métricas de agrupamento do algoritmo K-Means, no conjunto de dados ImageNet,

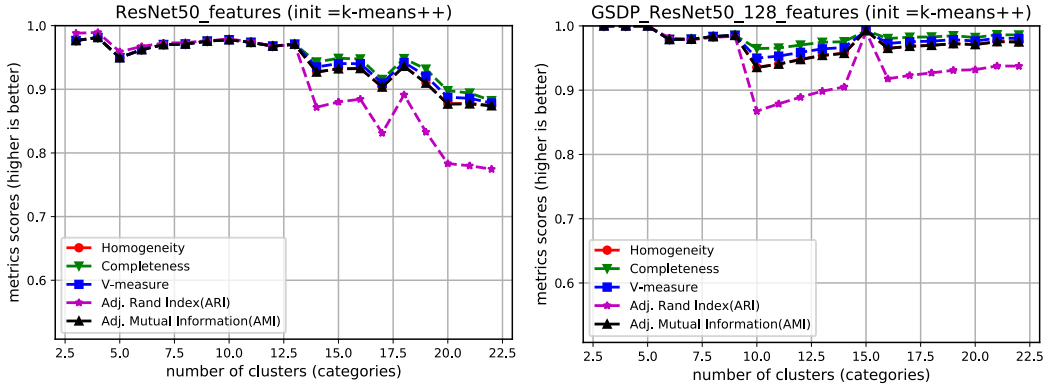
Descriptor	Size	FPS	Metrics Scores				
			H	C	V	ARI	AMI
Handcraft Features Performance on ImageNet (Russakovsky et al., 2015a)							
GIST (Oliva & Torralba, 2001)	960	0.82	0.05	0.05	0.05	0.01	0.05
LBP (Ojala et al., 2002)	512	0.72	0.02	0.03	0.03	0.01	0.02
HOG (Dalal & Triggs, 2005a)	1960	33	0.04	0.04	0.04	0.01	0.03
Color64 (Li, 2007)	64	8	0.12	0.12	0.12	0.04	0.11
Color_Hist(Song et al., 2004)	512	26	0.08	0.08	0.08	0.03	0.07
Hu_H_CH (Haralick et al., 1973)	532	6.9	0.04	0.04	0.04	0.01	0.02
Deep Features Performance on ImageNet (Russakovsky et al., 2015a)							
VGG16 (Simonyan & Zisserman, 2014)	4096	15	0,87	0,88	0,88	0,78	0,87
VGG_PCA_256	256	12.5	0,89	0,90	0,89	0,82	0,89
VGG_PCA_1024	1024	12.5	0,89	0,89	0,89	0,81	0,89
GSDP_VGG_256 (our)	256	12.8	**0,97**	**0,99**	**0,98**	**0,93**	**0,97**
GSDP_VGG_1024 (our)	1024	11.6	0,94	0,98	0,96	0,84	0,94
ResNet50 (He et al., 2016)	2048	10.6	0,88	0,90	0,89	0,78	0,88
ResNet50_PCA_128	128	12.5	0,88	0,88	0,88	0,81	0,88
ResNet50_PCA_512	512	12.5	0,89	0,90	0,90	0,82	0,89
GSDP_RNet_128 (our)	128	9.6	**0,97**	**0,98**	**0,98**	**0,93**	**0,97**
GSDP_RNet_512 (our)	512	9	0,91	0,97	0,94	0,73	0,91
Deep Features Performance on Coco (Lin et al., 2014)(CrossDataset)							
VGG16 (Simonyan & Zisserman, 2014)	4096	15	0.32	0.34	0.33	0.15	0.31
VGG_PCA_256	256	12.5	0.35	0.37	0.36	0.19	0.34
VGG_PCA_1024	1024	12.5	0.35	0.37	0.36	0.18	0.34
GSDP_VGG_256 (our)	256	12.8	**0.47**	**0.72**	**0.57**	**0.23**	**0.56**
GSDP_VGG_1024 (our)	1024	11.6	0.46	0.54	0.49	0.17	0.49
ResNet50 (He et al., 2016)	2048	10.6	0.29	0.36	0.32	0.17	0.31
ResNet50_PCA_128	128	12.5	0.32	0.34	0.33	0.17	0.31
ResNet50_PCA_512	512	12.5	0.34	0.35	0.34	0.20	0.33
GSDP_RNet_128 (our)	128	9.6	**0.43**	**0.69**	**0.53**	**0.26**	**0.52**
GSDP_RNet_512 (our)	512	9	0.34	0.47	0.40	0.09	0.39

Tabela 6.2: Métricas de agrupamento alcançadas pelas representações de cada descritor global selecionado. Mostram-se os valores das métricas de agrupamento alcançadas pelas características de cada representação na iteração 20 do experimento (22 primeiras categorias de cada banco de dados usado nos experimentos). Apresenta-se em negrito o melhor desempenho. Legenda: *Homogeneity* (H), Completeness (C), *V-measure* (V), *Adjusted Rand Index* (ARI) e *Adjusted Mutual Information* (AMI).

quando o número de partições (categorias) aumenta em cada execução do algoritmo. Apresenta-se o histórico das métricas de agrupamento do algoritmo K-Means nas primeiras categorias do banco de dados ImageNet para características construídas com os modelo VGG16, ResNet50, e o descritor GSDP proposto. O desempenho das métricas de agrupamento foram analisadas para vários métodos de inicialização do algoritmo K-Means: o método de inicialização padrão (*k-means++*) que acelera a convergência selecionando de forma inteligente os centros dos *clusters* iniciais; a inicialização aleató-



(a)

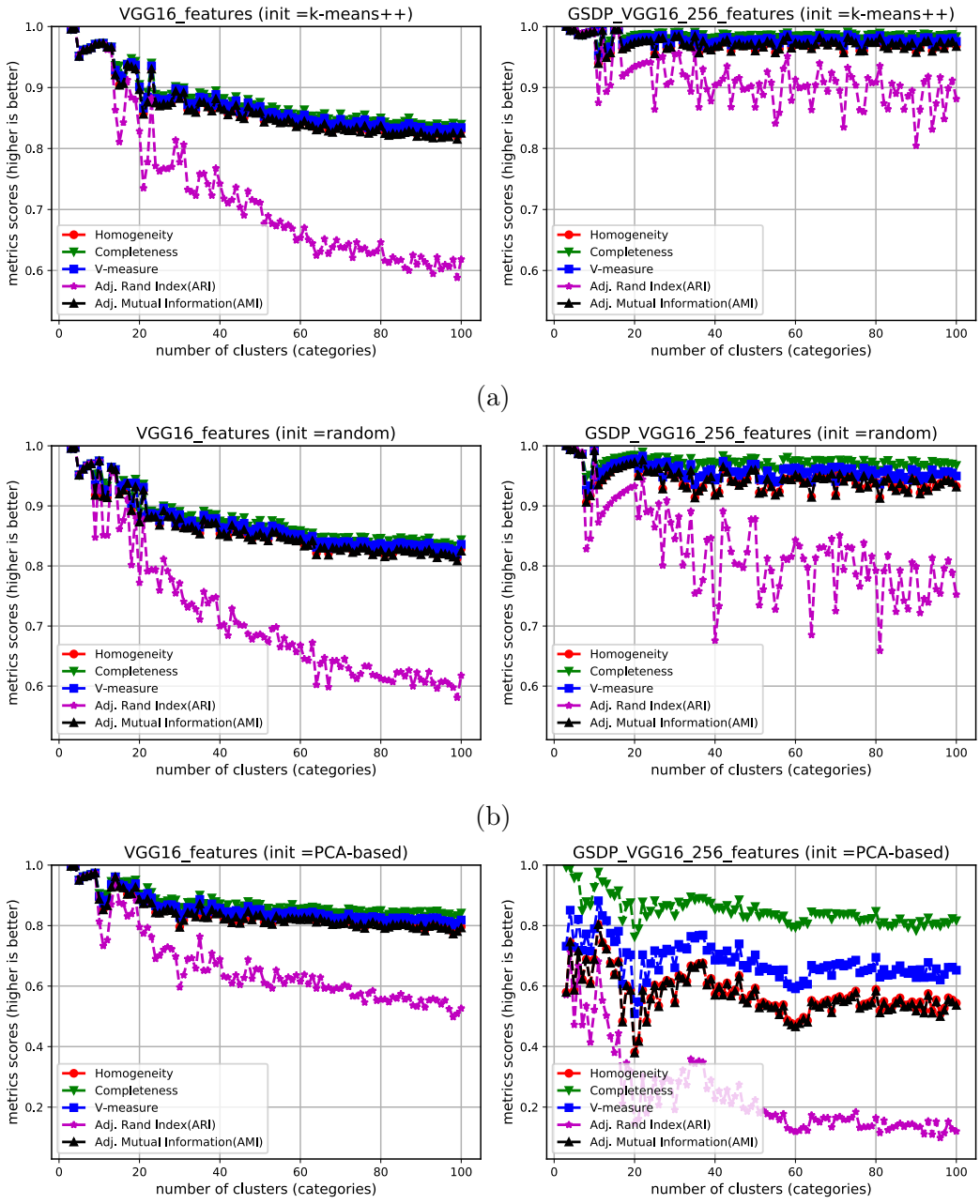


(b)

Figura 6.9: *Análise do desempenho em tarefas de agrupamento da codificação semântica proposta para as primeiras 20 iterações do algoritmo K-Means no banco de dados ImageNet.* Apresenta-se o histórico das métricas de agrupamento do algoritmo K-Means nas primeiras 22 categorias do banco de dados ImageNet para características extraídas com os modelos VGG16 e ResNet50 (Na esquerda) e características construídas com o descritor GSDP usando o modelo correspondente (Na direita). (a) Comparação das métricas K-Means para características construídas usando o modelo VGG16. (b) Comparação das métricas K-Means para características construídas usando o modelo ResNet50.

ria (*random*) que escolhe como centroides iniciais observações aleatórias dos dados; e a inicialização (*PCA-based*) que primeiro aplica o algoritmo PCA para reduzir o número de características antes de executar o agrupamento.

A Figura 6.9 apresenta o histórico do comportamento das métricas do algoritmo K-Means no agrupamento das primeiras 20 categorias das assinaturas construídas com os modelos VGG16 e GSDP no banco de dados ImageNet. Observa-se em cada experimento como, na medida que aumenta o número de partições (e os dados), deterioram-se as métricas de agrupamento, mas a abordagem de descrição semântica proposta (GSDP) consegue construir representações suficientemente discriminativas



(a)

(b)

(c)

Figura 6.10: *Análise do desempenho em tarefas de agrupamento da codificação semântica proposta para as primeiras 100 categorias do banco de dados ImageNet.* Apresenta-se o histórico das métricas de agrupamento do algoritmo K-Means nas primeiras 100 categorias do banco de dados ImageNet para características construídas com o modelo VGG16 (Na esquerda) e com o descritor GSDP proposto (Na direita). (a) Algoritmo K-Means com inicialização padrão (*k-means++*). (b) Algoritmo K-Means com inicialização aleatória (*random*). (c) Algoritmo K-Means com inicialização baseada em PCA (*PCA-based*).

como para ter maior robustez ante o aumento de variação da informação que produz o aumento das categorias de agrupamento. Por exemplo, a Tabela **??** mostra a captura realizada ao histórico das métricas de agrupamento K-Means apresentado na Figura 6.10a para o agrupamento de 20 partições (20 categorias). Observa-se como nesse estado dos experimentos, todos os valores das métricas de agrupamento K-Means relacionadas com as características-VGG16 são inferiores a 0.8; enquanto a maioria das métricas de agrupamento alcançadas pelas representações da codificação GSDP proposta permanecem acima de 0.9.

A Figura 6.10 apresenta o histórico do comportamento das métricas do algoritmo K-Means no agrupamento de características construídas com o modelo VGG16, mas nesse caso os experimentos foram estendidos até as primeiras 100 categorias do banco de dados ImageNet. Observa-se como no comportamento observado nas Figuras 6.10a e 6.10b todas as métricas de agrupamento K-Means caem acentuadamente para menos de 0.85 no agrupamento das características-VGG16, enquanto a maioria das métricas de agrupamento das representações-GSDP permanecem acima de 0.9.

A Figura 6.10c mostra o comportamento das métricas de agrupamento K-Means quando o algoritmo usa a inicialização PCA. Nesse caso, o histórico das métricas de agrupamento das representações-GSDP mostram um comportamento desordenado e pouco robusto. Mesmo quando esse comportamento não constitui um bom desempenho –em comparação com o comportamento observado para outras inicializações do K-Means (*k-means++*, *random*)– constitui um resultado esperado. O procedimento matemático PCA reduz a dimensionalidade da característica representando todos os n vetores de dados como combinações lineares de um pequeno número de autovetores, enquanto minimiza o erro médio quadrático e preserva a variância dos dados.

O comportamento das métricas de agrupamento observado na Figura 6.10c justifica-se pela natureza da representação semântica das características do descritor GSDP proposto. Por construção, as características unitárias posicionadas na segunda metade da representação semântica proposta (GSDP) estão altamente correlacionadas com as características unitárias posicionadas na primeira metade da assinatura-GSDP. Na estrutura da representação semântica proposta, a *diferença semântica* é construída a partir da diferença existente entre o *significado semântico do objeto* e o *protótipo semântico da categoria* (Ver Figura 6.1e)). Ou seja, na correlação existente entre as características unitárias da representação proposta, o protótipo da categoria possui um papel protagonista. Consequentemente, a inicialização PCA do agrupamento K-Means afeta o desempenho –em termos de métricas de agrupamento– da codificação proposta.

Os resultados apresentados na Figura 6.10c mostram que o descarte de características unitárias do objeto no domínio semântico pode ser problemático (*intensio-*

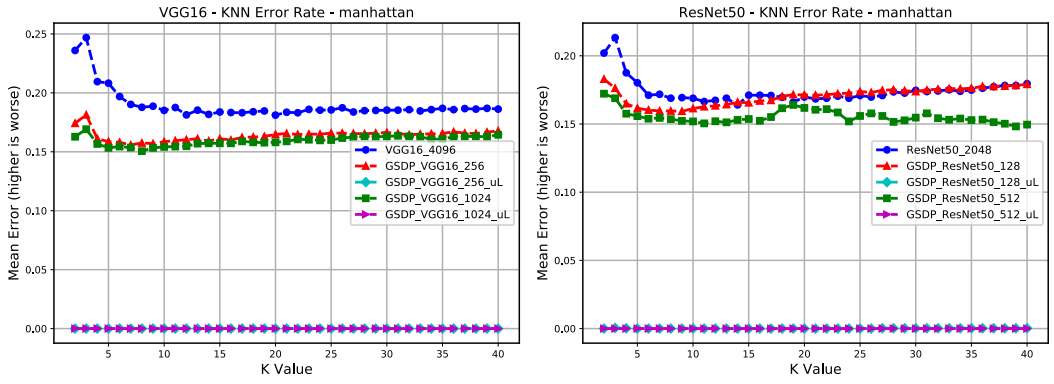
nal non-discreteness), pois o significado das características unitárias que compõe as representações semânticas é desconhecido a priori. Por exemplo, algumas características unitárias da representação proposta não representam características observadas do objeto, ao contrário, representam a diferença semântica do objeto com relação ao protótipo da categoria. Observa-se que, mesmo nessas condições desfavoráveis, a representação semântica GSDP consegue um desempenho razoável (ou melhor) –em termos de métricas de agrupamento– com relação às outras abordagens de descrição avaliadas.

Os resultados mostram que o descritor GSDP proposto constrói representações de objetos mais discriminativas e que superam, significativamente, às outras codificações globais avaliadas em termos de métricas de agrupamento. Esse resultado pode ser consequência de que a abordagem proposta visa construir famílias de representações de objetos baseadas nos protótipos das categorias. Ou seja, o protótipo semântico da categoria rege as representações construídas para cada membro da categoria, fato que pode ter influência no desempenho do agrupamento de imagens. A representação GSDP proposta preserva as informações semânticas contidas nas características extraídas com os modelos-CNN em uma representação semântica mais discriminativa e com uma dimensionalidade ainda menor. O desempenho alcançado pela codificação semântica proposta na tarefa de agrupamento de imagens motiva avaliar a capacidade de generalização da representação semântica GSDP em outras tarefas de Visão Computacional como a classificação de imagens.

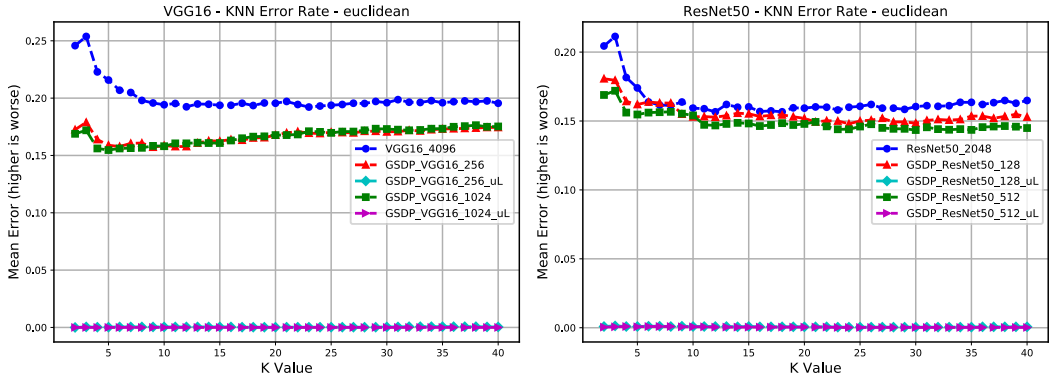
Classificação

Por construção, o descritor GSDP proposto constrói as representações semânticas das imagens de objetos com base na predição de classificação realizada pelo modelo CNN usado como modelo-base (Vide Figura 6.1b e Algoritmo 2 linha 4). Consequentemente, um erro de predição do classificador gera que o descritor GSDP construa uma representação semântica da imagem do objeto usando um protótipo semântico que não corresponde à categoria do objeto. Esse comportamento da abordagem de descrição proposta não é problemático se é considerado que os seres humanos descreverão erroneamente um objeto se o objeto for previamente reconhecido erradamente.

Nos experimentos realizados avaliou-se o desempenho da representação GSDP proposta em tarefas de classificação de imagens. Os experimentos foram realizados com representações GSDP de imagens de objetos construídas considerando dois cenários diferentes: *i)* assinaturas GSDP construídas com base na predição do modelo-CNN base (comportamento padrão do descritor GSDP); *ii)* assinaturas GSDP construídas com base na predição de um modelo de classificação ideal (100% de acurácia)(um com-



(a)



(b)

Figura 6.11: Taxa de erro alcançada, na tarefa de classificação KNN, por cada representação avaliada nas primeiras 100 categorias do banco de dados ImageNet. Variou-se o parâmetro K do algoritmo KNN para comparar, na tarefa de classificação de imagens, o desempenho das características VGG16 e ResNet50 *versus* as assinaturas GSDP correspondentes. Usou-se como medida de similaridade entre as características: *a)* distância de Manhattan; *b)* distância Euclidiana. O comprimento de cada característica foi especificado na legenda correspondente.

portamento hipotético do descritor GSDP). Usou-se o rótulo da categoria de imagem, anotado no banco de imagens ImageNet, como a predição do modelo de classificação ideal. O experimento realizou-se com o objetivo de analisar o possível desempenho da representação GSDP se o erro de seleção do protótipo semântico for zero (erro de predição do modelo CNN usado -em segundo plano- como extrator de características).

Nos experimentos foi utilizado o algoritmo KNN pois, semelhante ao algoritmo t-SNE, as representações das imagens são classificadas de acordo com a vizinhança local entre as características. Também analisou-se o desempenho da representação GSDP aumentando o valor do parâmetro do algoritmo KNN (K - número de vizinhos) e usando as distâncias Euclidiana e Manhattan como medidas de similaridade entre as características. Nesse experimento usou-se a mesma amostra de imagens do banco de dados

ImageNet usada para a avaliação das assinaturas GSDP na tarefa de agrupamento de imagens.

A Figura 6.11 mostra um exemplo do comportamento, na tarefa de classificação usando o algoritmo KNN, das características dos modelos VGG16 e ResNet50 em comparação com as assinaturas-GSDP correspondentes. Os experimentos foram realizados usando representações GSDP construídas nos dois cenários anteriormente mencionados (as representações GSDP construídas usando os rótulos da categoria como predição de classificação foram mostrados com o sufixo $\_uL$ na Figura 6.11).

Os resultados mostraram que a codificação GSDP proposta consegue superar o desempenho das codificações VGG16 e ResNet50 na tarefa de classificação KNN de imagens de objetos. Também, os experimentos mostraram que a representação GSDP, usando o modelo ResNet50, atingiu um desempenho melhor que aquelas representações construídas usando o modelo VGG16. Além disso, observou-se que as representações GSDP construídas usando como predição de classificação os rótulos da categoria (rotuladas com _ul na Figura 6.11) são altamente discriminativas (erro médio próximo de 0.1). Consequentemente, pode-se concluir que a codificação semântica de imagens de objeto proposta consegue melhorar substancialmente seu desempenho na medida que aumenta a precisão do modelo-CNN de classificação usado como modelo-base.

6.5 Discussão

O presente capítulo apresenta a abordagem de *descrição semântica global baseada em protótipos* proposta. O descritor GSDP proposto introduz a semântica, encapsulada e interpretada com o modelo CPM proposto, na descrição semântica global de imagens de objetos. A metodologia de descrição semântica apresentada visa construir uma codificação das características da imagem do objeto que encapsula o significado do objeto no contexto da categoria a qual pertence. Essas representações e interpretações da imagem são construídas mediante os protótipos semânticos calculados com o conhecimento aprendido pelos modelos CNN de classificação pré-treinados.

Diferente das abordagens existentes na literatura, o enfoque proposto introduz uma nova perspectiva de descrição semântica que visa simular a estratégia humana de descrever objetos destacando as diferenças que os tornam semanticamente distintivos dentro da categoria. O método proposto descreve semanticamente o objeto baseado na comparação de todas as características unitárias que o compõem com relação a todas as características representativas da categoria (características encapsuladas no protótipo semântico).

O método de descrição semântica proposto baseia-se nas características – de imagens de objetos– extraídas pelos modelos CNN de classificação. Essa particularidade constitui um pré-requisito estabelecido nesta pesquisa na tentativa de simular o comportamento humano de descrever os objetos baseado nas mesmas características aprendidas para classificá-los. Consequentemente, o modelo CNN de descrição semântica proposto possui a mesma arquitetura que os modelos CNN de classificação e usa o mesmo conhecimento aprendido das categorias de objetos. Ou seja, o modelo de descrição semântica de objetos proposto não precisa ser treinado novamente para a aprendizagem da representação semântica do objeto, pois somente precisa das características-CNN extraídas do objeto (características aprendidas pelos modelos CNN de classificação). A abordagem proposta possui a vantagem de que o surgimento de novos modelos de classificação com maior acurácia beneficiaria também o desempenho da abordagem de descrição semântica proposta (aumentaria o desempenho). Adaptar o descritor semântico GSDP proposto a um novo modelo CNN de classificação é uma tarefa simples, basta calcular apenas os protótipos semânticos correspondentes ao novo modelo CNN de classificação.

Diferente de outras abordagens, a codificação semântica proposta permite que as assinaturas semânticas do descritor GSDP proposto possam ser interpretadas. As propriedades do descritor semântico GSDP permitem preservar nas assinaturas construídas: *i)* o *significado semântico do objeto* usado para a classificação nos modelos-CNN; e *ii)* a *distância prototípica* do objeto na categoria, a qual pode ser entendida como uma pontuação da tipicidade do objeto dentro da categoria. Os experimentos realizados mostraram que essas propriedades fazem possível, similar ao apresentado no Capítulo 5, usar a assinatura do descritor para representar o objeto na posição semântica correspondente a sua tipicidade dentro da categoria. Ou seja, a representação semântica proposta encapsula a informação semântica necessária do objeto, pelo modelo CPM proposto, para simular a organização prototípica dos elementos das categorias de objetos.

Os experimentos mostraram que a abordagem proposta para a descrição semântica global de imagens de objetos pode superar outras representações globais da imagem em várias tarefas de visão computacional. Os experimentos mostraram que as representações GSDP de dimensões mais baixas (para cada modelo CNN avaliado) foram as que alcançaram o melhor equilíbrio entre tamanho e desempenho.

Capítulo 7

Classificação semântica baseada em Protótipos

Vários trabalhos (Rosch, 1977; Estes, 1986; Rosch, 1978; Tulving, 2007; Martin, 2007; Collins & Curby, 2013) mostraram que a memória semântica do ser humano envolve a definição semântica (ou significado semântico) dos objetos. Esses trabalhos concluíram que o sucesso das tarefas de reconhecimento e classificação de objetos está altamente e causalmente relacionado com o sucesso da tarefa de recuperação desse conhecimento semântico aprendido.

Um objetivo constante das áreas de Visão Artificial e Aprendizagem de Máquina constitui o desenvolvimento e aperfeiçoamento de modelos de classificação complexos que possuam um desempenho similar ao ser humano. Durante anos, os modelos de classificação e detecção de objetos foram baseados em modelos discriminativos como *Nearest Neighbor* (Altman, 1992), *Support Vector Machines* (Cortes & Vapnik, 1995), *Boosting* (Breiman, 1996), Classificadores Lineares (Ng & Jordan, 2002), etc. Entretanto, nos últimos anos, o retorno das CNNs melhorou o desempenho das tarefas de classificação de imagens de objetos em níveis sem precedentes (Simonyan & Zisserman, 2014; Chollet, 2016; He et al., 2016; Szegedy et al., 2016; Howard et al., 2017; Szegedy et al., 2017). As CNNs abriram a possibilidade - pela primeira vez - de obter um modelo computacional de reconhecimento visual com comportamento similar à memória semântica em tarefas de classificação de imagens a grande escala. Além disso, os modelos CNN demonstraram alta capacidade de generalização para lidar com dados de alta variabilidade semântica.

Apesar da grande variedade de modelos CNN, da diversidade de suas aplicações em tarefas de processamento semântico da imagem (Karpathy & Fei-Fei, 2015; Yi et al., 2016; Lin et al., 2016a; Nogueira et al., 2017), e do salto qualitativo alcançado nos úl-

timos anos, o fundamento teórico do poder interpretativo dos modelos CNN continua sendo limitado. Os resultados alcançados pelo paradigma das redes neurais profundas propiciaram que alguns estudos cognitivos e da neurociência (Yamins et al., 2014; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Cichy et al., 2017) pesquisaram as ligações existentes entre as redes neurais profundas e o sistema visual no cérebro humano. Esses autores concluíram que a combinação ponderada das características da última camada totalmente conectada dos modelos CNN, pode explicar *completamente* o córtex temporal inferior dos humanos.

Nesse ponto, onde está a diferença entre os modelos CNN e o modelo de reconhecimento visual humano? Alguns trabalhos mostraram algumas diferenças existentes no comportamento de ambos os modelos. Szegedy et al. (2013) mostraram que é fácil construir imagens contraditórias que enganam aos modelos CNN, mas que são facilmente reconhecidas pelo sistema visual humano. Nguyen et al. (2015) geraram imagens que são completamente irreconhecíveis pelos humanos, mas que um modelo CNN consegue classificar com $99,99\%$ de confiança. Essa estratégia para enganar os modelos CNNs levanta questões sobre as verdadeiras capacidades de generalização e interpretação visual de tais modelos.

Outras diferenças relevantes existem desde o ponto de vista conceitual dos modelos: os modelos CNN são treinados estritamente para aprender parâmetros que conseguem otimizar a capacidade de predição dos rótulos da categoria; em oposição à predição ou aprendizagem de características ausentes nas categorias ou à criação de um modelo generativo de dados. Mas, a diferença mais visível reside na tomada de decisão (*decision making*) após a extração de características: a maioria dos modelos de classificação CNN usam a função softmax para avaliar as desejabilidades de escolhas alternativas e selecionar uma categoria particular. A função *softmax* realiza a tomada de decisão baseada na ativação construída com as características da camada anterior (o valor semântico definido nesta pesquisa), mas não existe uma interpretação real do significado semântico que possui esse valor de ativação. Assim, mesmo quando a função softmax consegue modelar eficientemente alguns processos cognitivos (Lee & Wang, 2009), ainda se pesquisa como modelar a tomada de decisões realizada pelo cérebro humano.

Uma teoria que justifica como os seres humanos realizam a tomada de decisões nas tarefas de classificação de objetos vêm dos estudos semânticos cognitivos relacionados com a *Teoria dos Protótipos* (Rosch, 1975b; Rosch & Mervis, 1975). Assim, de acordo com o *modelo do protótipo* (Homa & Vosburgh, 1976; Posner & Keele, 1968; Reed, 1972; Zaki et al., 2003), os seres humanos representam os protótipos das categorias como a tendência central calculada das instâncias de objetos da categoria (por exemplo, na

Figura 1.1a dos elementos apresentados em 1) o homem armazena 3) como significado principal), e classificam os objetos com base em quão similares são com os protótipos das categorias aprendidas (Ver Figura 1.1b).

A metodologia de descrição global semântica de objetos apresentada no Capítulo 6 não categoriza o objeto de entrada baseado na similaridade de suas características com relação ao protótipo semântico da categoria. Ou seja, o protótipo da categoria não está incluído no processo de categorização. O descritor GSDP proposto reconhece (*recupera*) o protótipo da categoria usando uma abordagem bem simples: primeiramente classifica a imagem (objeto) de entrada usando um modelo CNN pré-treinado (Ver Figura 6.1b), e seguidamente seleciona (com o índice da categoria) o protótipo correspondente no banco de dados de protótipos semânticos construídos (Ver Figura 6.1c-d). Assim, o descritor semântico GSDP proposto usa o modelo CPM para descrever semanticamente o objeto usando o protótipo da categoria; mas não usa o o modelo CPM no processo de categorização do objeto.

Neste Capítulo apresenta-se uma abordagem para *recuperar* o protótipo da categoria baseado na sua similaridade com a informação visual da imagem de objeto de entrada. Propõe-se uma metodologia para introduzir o conceito de *categorização baseado em protótipos* nos modelos CNN de classificação. O método de classificação que se apresenta procura usar a semântica capturada pelo modelo CPM proposto na camada totalmente conectada dos modelos CNN de classificação. A metodologia proposta visa substituir a tomada de decisão da classificação baseada na camada de ativações *softmax*, pela decisão de classificar os objetos com base em quão semelhantes são com os protótipos semânticos das categorias de objetos.

7.1 Influência da tipicidade na classificação

Rosch (1975b, 1978) obteve evidências de que os humanos armazenam o *significado semântico da categoria* baseado nos graus de representatividade (ou tipicidade) dos membros da categoria. Os resultados dos experimentos de Rosch (Rosch, 1975b, 1978) mostraram que os seres humanos conseguem melhorar o desempenho da classificação de novos exemplares de objetos quando aprendem as categorias correspondentes usando membros típicos das categorias. Assim, a autora mostrou que a abordagem humana de representar o conhecimento das categorias de objetos mediante a *organização prototípica* dos membros possui influência no processamento on-line (Rosch, 1978). Segundo Rosch (1975b) a organização prototípica da categoria é uma representação semântica da categoria baseada na tipicidade do objeto, onde objetos típicos constituem o centro

semântico da categoria (o protótipo), e os demais membros da categoria são posicionados próximos ou distantes desse centro semântico da categoria baseado no nível de tipicidade (representatividade) de cada elemento dentro da categoria.

Saleh et al. (2016) baseou-se nos resultados e conclusões dos experimentos de Rosch sobre o processo de aprendizagem baseado em tipicidade visual, para introduzir esse concepto de tipicidade visual no método de aprendizagem dos modelos CNN de classificação de imagens. Os autores propuseram uma medida de tipicidade visual da imagem baseada na probabilidade de predição de classificadores SVM construídos para 6 categorias de objetos (*one-class SVM*). Saleh et al. (2016) introduziram essa medida de tipicidade/atipicidade das amostras de treinamento como um termo de ponderação na função de perda dos modelos CNN. Assim, na abordagem usada pelos autores, o modelo CNN necessita de um modelo SVM de classificação para cada categoria de objeto, e cada modelo é usado como o estimador da tipicidade do objeto. A abordagem de Saleh et al. (2016) possui a desvantagem do alto custo computacional, o que impede que seja facilmente generalizado a qualquer arquitetura de modelos CNN de classificação. Mas, o resultado mais relevante desse trabalho é que os autores mostraram que a aprendizagem com maior ênfase em amostras representativas (típicas) aumenta a capacidade de generalização dos classificadores CNN treinados de maneira discriminativa.

Baseado nesses supostos, neste capítulo propõe-se um método de aprendizagem que propõe classificar o objeto baseado na relevância ou tipicidade visual do objeto dentro da categoria. No Capítulo 5 foi mostrado que o modelo CPM proposto consegue capturar a tipicidade do objeto, e que esse comportamento prototípico dos elementos dentro da categoria está relacionado com a representação do *protótipo semântico* e a *distância prototípica* proposta. Doravante apresenta-se como introduzir a semântica da imagem do objeto capturada pelo modelo CPM proposto nos modelos CNN de classificação.

7.2 Introduzindo o protótipo na classificação

O conceito de categorização baseado em protótipos é uma teoria que entende a tarefa de reconhecimento de objetos como uma tarefa de recuperação de abstrações das categorias de objetos armazenadas na memória. A escolha de como é construído a representação dessas abstrações (os protótipos) influencia fortemente no desempenho da categorização baseada em protótipos (Wohlhart et al., 2013).

A maioria dos trabalhos que tentaram introduzir o protótipo da categoria na ta-

refa de categorização (Crammer et al., 2003; Seo & Obermayer, 2003; Wohlhart et al., 2013; Snell et al., 2017), construíram modelos com arquiteturas próprias que usavam como heurística - ou método de aprendizagem - aprender (ou otimizar) uma representação do protótipo que tentava deslocar o protótipo para posições próximas às amostras de treinamento da mesma categoria, mas distante das amostras de outras categorias. Nessas abordagens o processo de calcular os protótipos é custoso porque os protótipos formam parte das funções de aprendizagem que propagam o error de classificação através da rede (*Backpropagation*). Esses modelos assumem que as características unitárias das amostras possuem a mesma importância dentro da categoria pelo que as normas L1 e L2 são as métricas usadas para calcular a distância entre os elementos e o protótipo correspondente. Outra característica desses trabalhos é que a tipicidade das amostras da categoria não é considerada relevante para a método de aprendizagem do modelo de classificação e, consequentemente, os protótipos são calculados tendo em conta todos os elementos da categoria.

Aliás, Bendale & Boult (2016) e Snell et al. (2017) propuseram trabalhos que usam como o centroide da categoria o vetor médio calculado com as características extraídas de todos os membros da categoria. Essa abordagem assume que todos os objetos são igual de representativos para a categoria, pelo que agregar novos elementos (representativos ou não representativos) à categoria sempre gera mudanças no centroide da categoria.

7.2.1 Características da abordagem proposta

A metodologia que se apresenta possui diferenças significativas com respeito aos trabalhos de classificação baseada em protótipos anteriormente apresentados (Crammer et al., 2003; Seo & Obermayer, 2003; Wohlhart et al., 2013; Bendale & Boult, 2016; Snell et al., 2017). O método de classificação proposto usa os conceitos principais da Teoria dos Protótipos através do modelo CPM proposto. A abordagem proposta distingue-se por:

- A representação do protótipo semântico. A representação do protótipo é baseada na hipótese de que os seres humanos podem aprender melhor as categorias de objetos observando apenas as amostras típicas (Rosch, 1975b, 1978). Consequentemente, a codificação de protótipo semântico é calculada usando apenas as amostras de imagens típicas. Contrário à abordagem de usar o vetor médio como centroide da categoria, a representação proposta possui a vantagem de que o centro semântico da categoria é mais invariante às mudanças, pois esse centroide só mudará quando os elementos agregados à categoria sejam, especifi-

camente, elementos representativos (típicos) da categoria. Além disso, calcular a representação proposta requer menos dados;

- A função de similaridade. A medida de similaridade proposta baseia-se em algumas medidas de similaridades psicológicas. A metodologia proposta usa, no processo de aprendizagem, a distância prototípica proposta como métrica de distância semântica entre os elementos. A métrica de distância semântica proposta pode ser entendida como a pontuação de tipicidade do objeto, consequentemente, a ideia principal é introduzir a representatividade visual da imagem no processo de aprendizagem da categoria;

- A simplicidade e escalabilidade. A representação do protótipo proposta não depende da aprendizagem dos pesos da rede CNN (*Backpropagation*), consequentemente a abordagem proposta é menos complexa que outros trabalhos na literatura. A camada PS-Layer é fácil de usar e converte um modelo CNN comum em um modelo CNN de classificação baseado em protótipos sem fazer mudanças substanciais na arquitetura do modelo CNN original;

- A interpretabilidade. A maioria dos modelos de classificação CNN usam a função softmax (sobre o valor semântico) como métrica para a tomada de decisões. Mesmo quando essa abordagem pode modelar com eficiência alguns processos cognitivos (Lee & Wang, 2009) e obter uma alta precisão na classificação de imagens, ainda não se sabe como fazer uma interpretação precisa dos resultados dos modelos CNN de classificação. Por outro lado, a camada PS-Layer proposta fornece maior poder interpretativo aos modelos CNN devido à simplicidade e fácil interpretação geométrica do conceito de tipicidade da imagem do objeto (Ver Figura 5.1).

Em resumo, similar ao método de aprendizagem humana, o modelo de classificação que se propõe pretende reger a aprendizagem do modelo CNN mediante a simulação da organização prototípica da estrutura interna das categorias. Para atingir esse objetivo, a metodologia proposta assume que é possível classificar os objetos baseado na similaridade dos elementos com relação à representação do protótipo semântico proposto. Observa-se que essa abordagem de classificação tenta classificar o objeto baseado na sua representatividade semântica dentro da categoria. Com esse fim, o método proposto precisa como pré-requisitos: *i)* um modelo CNN pré-treinado para calcular os protótipos semânticos; *ii)* conhecer a priori todos os protótipos semânticos das categorias. A abordagem proposta mantém fixa a posição do protótipo da categoria (previamente calculado) enquanto aprende reestruturar o domínio m-dimensional

das características dos objetos para conseguir a organização prototípica da categoria. Essa estratégia visa alcançar uma nova representação das características do objeto que posiciona os elementos representativos próximos do protótipo semântico da categoria, e outros menos representativos, mais distantes.

7.3 Camada de Similaridade Prototípica

Sob a perspectiva das premissas desta pesquisa, formula-se a hipótese de que é possível melhorar o desempenho de modelos CNN de classificação pré-treinados treinando-os novamente com os protótipos semânticos calculados a partir de membros típicos das categorias de objetos. Com o objetivo de introduzir o *conceito de categorização baseada em protótipos* nos modelos CNN comuns, propõe-se uma camada de decisão semântica denominada Camada de Similaridade Prototípica (*Prototypical Similarity Layer (PS-Layer)*). A camada PS-Layer utiliza os *protótipos semânticos* calculados como *conhecimento a priori* em seus neurônios e utiliza a *distância prototípica* proposta como valor de ativação de cada neurônio.

Na Figura 7.1 apresenta-se a estrutura interna da camada PS-Layer proposta. A Figura 7.2 apresenta como pode ser integrada a camada convolucional proposta (PS-Layer) na metodologia geral desta pesquisa. Diferente da abordagem de Jetley et al. (2015), que usa uma imagem *template* como conhecimento a priori da informação do objeto, a camada PS-Layer usa os protótipos semânticos calculados (Vide o Algoritmo 1) como conhecimento a priori da informação semântica da categoria. A camada PS-Layer classifica o objeto considerando a sua similaridade com relação a esses protótipos semânticos calculados. Ou seja, cada i-ésimo neurônio da camada PS-Layer armazena na memória as informações semânticas representativas da i-ésima categoria de objeto que se encontram encapsuladas no protótipo semântico (P_i) correspondente (Ver Figura 7.1b). A saída do i-ésimo neurônio é uma função da distância prototípica do objeto, e esse valor de saída constitui a probabilidade de classificação do objeto na i-ésima categoria. Assim, a saída da camada PS-Layer pode ser interpretada como uma distribuição de probabilidade da similaridade semântica entre o objeto e os protótipos semânticos aprendidos. Observa-se que a camada PS-Layer usa a métrica de distância semântica proposta (distância prototípica) como ativação do neurônio, com o intuito de conseguir que a probabilidade de classificação no i-ésimo neurônio seja diretamente proporcional à tipicidade do objeto na i-ésima categoria.

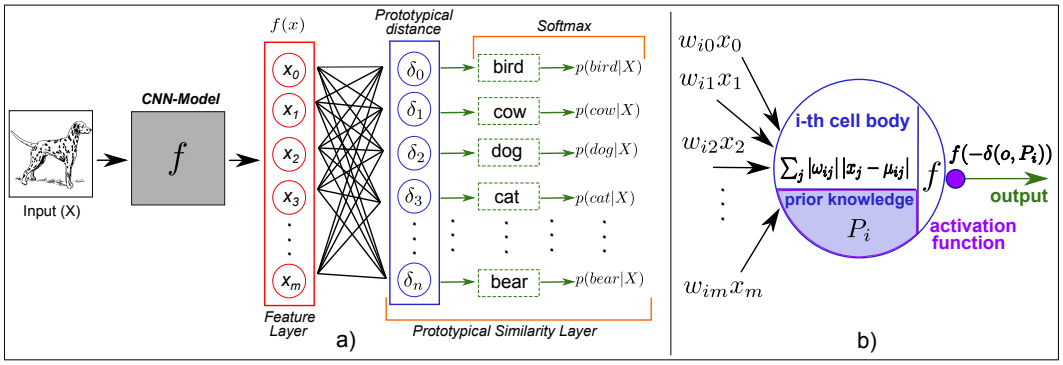


Figura 7.1: *Camada de Similaridade Prototípica proposta. a)* Exemplo de um modelo CNN que usa a camada PS-Layer; *b)* O modelo matemático do neurônio da camada PS-Layer. Observa-se que o corpo da célula mantém o protótipo da categoria como conhecimento a priori e usa a distância prototípica como ativação do neurônio. Fonte: Elaborado pelo autor.

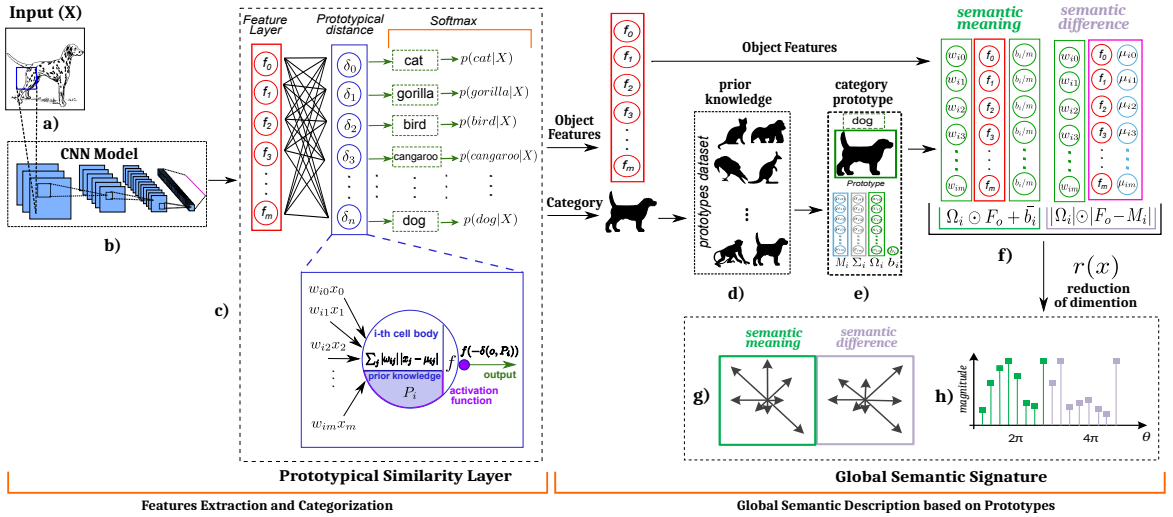


Figura 7.2: *Integração da camada PS-Layer na metodologia de descrição semântica proposta.* A metodologia pode ser dividida em duas etapas principais: *1)* Extração de características CNN e categorização; e *2)* Transformação das características CNN na assinatura GSDP proposta. *a)* imagem de entrada; *b)-c)* extração de características e classificação usando um modelo CNN de classificação previamente treinado. A camada PS-Layer proposta é usada para converter um modelo CNN comum em um modelo de classificação CNN baseado em protótipos; *d)-h)* processo de construção da assinatura GSDP proposta. Fonte: Elaborado pelo autor.

7.3.1 O modelo matemático do neurônio

A camada PS-Layer proposta visa introduzir as definições principais da Teoria dos Protótipos simuladas com o modelo CPM proposto (Ver Capítulo 5). Observa-se que cada i-ésimo neurônio, através do modelo CPM proposto, constrói a própria representação semântica da i-ésima categoria que representa. Isso é consequência de que o neurô-

nio conhece a priori, mediante o protótipo semântico da categoria, o centro semântico da categoria e os valores fronteiras das características representativas da categoria. Mas, por construção, o modelo CPM proposto não consegue definir os limites da categoria. Consequentemente, foram herdados dos modelos psicológicos formais GCM e MPM (Ver Seção 3.4) outros conceitos necessários para definir o modelo matemático do neurônio, e que permitam calcular a probabilidade de pertença do objeto à i-ésima categoria.

Similaridade Prototípica

A abordagem de classificação proposta fundamenta-se nos estudos do *modelo do protótipo* (Reed, 1972; Homa & Vosburgh, 1976; Minda & Smith, 2001) e visa usar a mesma função de *similaridade semântica* proposta pelos modelos formais GCM (Nosofsky, 1986) e MPM (Minda & Smith, 2001, 2002). Similar a esses modelos propõe-se que a *similaridade semântica* ($S(o, P_i)$) entre a imagem do objeto $o \in O$ e o protótipo da i-ésima categoria (P_i) está dada pela expressão:

$$S(o, P_i) = exp(-\alpha\delta(o, P_i)) \tag{7.1}$$

onde α é o mesmo parâmetro de *sensibilidade* da Equação 3.2, e $\delta(o, P_i)$ constitui a distância semântica entre o objeto $o \in O$ e o protótipo (P_i) proposta, $\forall i = 1...n$. É importante lembrar que a *distância prototípica* proposta $\delta(o, P_i)$ é não negativa por definição, pelo que usar os valores positivos do parâmetro α garante que a função exponencial da Equação 7.1 possua valores no domínio negativo do eixo x. Consequentemente, e pelas características da função exponencial, a *similaridade semântica* dos objetos possui valores entre $[0, 1]$ garantindo que elementos bem distantes do protótipo da categoria não sejam similares ($S(o, P_i) \approx 0$) e elementos bem próximos ao protótipo da categoria sim sejam reconhecidos como similares ao protótipo ($S(o, P_i) \approx 1$).

Penalidade prototípica

Observa-se que a *similaridade semântica* do objeto com o prototípico depende estritamente da distância prototípica e não de como está distribuído o vetor de características do objeto. Assim, a função de distância proposta deve garantir que os elementos com vetores de características marcadamente diferentes (por exemplo que não pertençam à mesma categoria) não possuam a mesma distância prototípica e, consequentemente, não seja atribuído o mesmo valor de similaridade semântica.

Não obstante, a *distância prototípica* proposta perde a distribuição espacial da característica do objeto quando o *vetor semântico residual* ($\vec{r} = |f_j - \mu_{ij}|$) é usado no produto escalar (Ver Equação 5.2). Consequentemente, um objeto específico $o \in O$ pode ter distâncias prototípicas semelhantes com protótipos de diferentes categorias ($\delta(o, P_i) \approx \delta(o, P_j), i \neq j$), mesmo quando as distribuições dos *vetores semânticos residuais* sejam notavelmente diferentes ($\vec{r_i} \neq \vec{r_j}$). Pode acontecer também que elementos de diferentes categorias ($o_1 \in O_{c_i}, o_2 \in O_{c_j}, i \neq j$) possam ter vetores residuais semelhantes no contexto de i-ésima categoria ($\vec{r_1} = \vec{r_2}$) gerando que os valores das distâncias prototípicas correspondentes sejam semelhantes.

Por essa razão, foi introduzida uma *penalidade* na métrica de distância proposta quando a distribuição espacial do *vetor semântico residual* ($\vec{r}$) é bastante diferente da distribuição que rege a fronteira do protótipo da i-*th* categoria ($\Sigma_i \in (M_i, \Sigma_i, \Omega_i)$). O propósito da *penalidade prototípica* (Ver Definição 11) é adicionar um valor de punição – à distância prototípica– que seja próximo a zero quando as distribuições das características comparadas sejam parecidas, mas o suficientemente grande quando são marcadamente diferentes.

Para construir esse valor de penalidade usaram-se os fundamentos principais da *Desigualdade de Chebyshev* ($\Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$) utilizados para definir as bordas do protótipo semântico (Ver Definição 7). A desigualdade de Chebyshev garante que não mais do que uma certa fração dos valores da amostra podem ser maiores que uma certa distância fixa da média da distribuição. Especificamente, não mais do que $\frac{1}{k^2}$ dos valores da distribuição podem ser maiores que k desvios-padrão (σ) da média (μ). Ou seja, é poco provável que o valor (X) de um elemento que pertença à distribuição se encontre a k desvios-padrão do valor esperado (μ). Consequentemente, elementos com valores (X) que cumprem a expressão $|X - \mu| - k\sigma \geq 0$ quando k é suficientemente grande, são improváveis que pertença à distribuição de valores que carateriza à categoria.

Definição 11. *Penalidade prototípica.* Seja $o \in O$ um objeto avaliado no contexto da i-ésima categoria $c_i \in C$, F_o as características do objeto o, e $P_i = (M_i, \Sigma_i, \Omega_i, b_i)$ o protótipo semântico da i-ésima categoria. Se o vetor de características F_o é marcadamente diferente do valor esperado das características da categoria (M_i), a distância prototípica do objeto será penalizada com o valor *penalidade prototípica* definido pela expressão *penalidade prototípica* $= \sum_{j=1}^{m} unitary\_penalty\__j(o, P_i)$ com :

$$unitary\_penalty\__j(o, P_i) = \begin{cases} |\omega_{ij}|\, u_j, & \text{if } u_j > 0 \\ 0, & \text{if } u_j \leq 0. \end{cases} \tag{7.2}$$

onde $u_j = (|f_j - \mu_{ij}| - \kappa\sigma_{ij}) \times \phi$, $\omega_{ij} \in \Omega_i$, $\mu_{ij} \in M_i$, $\sigma_{ij} \in \Sigma_i$ e $f_j \in F_o$; $\forall j = 1...m$; $\forall i = 1...n$. As constantes positivas $factor(\kappa)$ e $penalty(\phi)$ constituem hiperparâmetros (Adaptado da *Desigualdade de Chebyshev*).

Assim, a *distância prototípica penalizada* ($\delta_p(o, P_i)$) é a distância prototípica proposta multada com o valor da *penalidade prototípica*: $\delta_p(o, P_i) = \delta(o, P_i) + penalidade\ prototípica$.

Probabilidade prototípica

A camada *PS-Layer* proposta possui tantos neurônios quanto protótipos e categorias a serem reconhecidas (Ver Figura 7.1). A saída de cada i-ésimo neurônio usa as Equações 5.2 e 7.2 para calcular o valor de diferença semântica entre o objeto e o protótipo semântico correspondente à i-ésima categoria. Consequentemente, cada i-ésimo neurônio pode calcular a similaridade semântica (Vide Equação 7.1) do objeto com relação ao protótipo semântico da categoria que representa. Com essas informações relativas a cada i-ésimo neurônio, a camada *PS-Layer* consegue calcular a probabilidade de pertença do objeto $o \in O$ à i-ésima categoria.

Consequentemente, propôs-se uma generalização da equação de probabilidade do modelo psicológico formal MPM (Ver Equação 3.5) para calcular a distribuição de probabilidade que constitui a saída da camada *PS-Layer*. Sem perda de generalidade, e usando as mesmas assunções do modelo formal MPM, foi definido em 1 os valores dos parâmetros α e γ das Equações 3.5 e 7.1, respectivamente. Consequentemente, a camada PS-Layer proposta categoriza o objeto na i-ésima categoria com a probabilidade calculada através da expressão:

$$\begin{aligned} P(c_i|o) &= \frac{S(o, P_i)^\gamma}{\sum_{k=1}^{n} S(o, P_k)^\gamma} \\ &= \frac{exp(-\alpha\delta(o, P_i))^\gamma}{\sum_{k=1}^{n} exp(-\alpha\delta(o, P_k))^\gamma} \\ &= \frac{exp(-\delta(o, P_i))}{\sum_{k=1}^{n} exp(-\delta(o, P_k))} \end{aligned} \tag{7.3}$$

onde $\delta(o, P_k)$ é a *distância prototípica* entre o objeto $o \in O$ e o k-ésimo protótipo semântico (P_k) armazenado no k-ésimo neurônio da camada, $\forall k = 1...n$. Desse modo, a camada PS-Layer proposta transforma o vetor n-dimensional ($\vec{\delta}(o, P_k)$) composto pelas n distâncias prototípicas calculadas em cada i-ésimo neurônio, em um vetor normalizado que constitui uma distribuição de probabilidade composta pelas n-probabilidades

de classificação do objeto em cada uma das n categorias. Observa-se na Equação 7.3 que a distribuição de probabilidade de similaridades prototípicas da camada PS-Layer proposta pode ser facilmente construída usando a função *softmax* sobre os valores negativos do vetor n-dimensional ($\vec{\delta}(o, P_k)$) composto pelas n distâncias prototípicas calculadas ($P = softmax(-\vec{\delta}(o, P_k))$.

Gradiente do neurônio

Nesta seção apresentam-se alguns detalhes de construção da camada PS-Layer proposta. Especificamente, expõem-se passo a passo as principais diferenças e assunções introduzidas na modelagem da camada proposta para simplificar o cálculo do gradiente de cada neurônio com relação aos pesos aprendidos pela camada PS-Layer.

A função do neurônio em camadas CNN normalmente possui a forma: $neuron(x) = activation(g(x))$ onde x é a característica de entrada do neurônio (saída da camada anterior), $g(x)$ é uma transformação linear da entrada (comumente $g(x) = \omega x + b$), e $activation(g(x))$ é uma função que define o tipo de saída do neurônio, dada a entrada x (exemplo de funções de ativação: *logistic, tanh, RELU, softmax, etc*).

Em caso de modelos de classificação o neurônio da camada completamente conectada (*dense layer*) possui a forma: $neuron(x) = softmax(g(x))$. Usando a regra da cadeia para encontrar a derivada dessa composição de funções com relação aos pesos da camada, a derivada do neurônio é simplesmente:

$$\frac{\partial neuron}{\partial \omega} = \frac{\partial neuron}{\partial g(x)} \frac{\partial g(x)}{\partial \omega} = \frac{\partial softmax}{\partial g(x)} \frac{\partial g(x)}{\partial \omega}. \tag{7.4}$$

Observa-se que o primeiro fator da Equação 7.4 sempre terá a mesma derivada sem importar o tipo de função $g(x)$, consequentemente a expressão da derivada depende da derivada parcial da função de transformação da entrada $g(x)$ (por exemplo quando $g(x) = \omega x + b$, $\frac{\partial g(x)}{\partial \omega} = x$).

Similarmente, a camada PS-Layer proposta usa a expressão 7.4 para calcular o gradiente do i-ésimo neurônio, mas usando como transformação da entrada x a distância prototípica do objeto $g(x) = -\delta(x, P_i) = |\omega|\,|x - \mu_i|$. Sem perda de generalidade, foram introduzidas algumas restrições no método de aprendizagem dos pesos (ω) do neurônio para conseguir uma maior simplicidade na modelagem: *i)* os pesos devem ser não negativos ($\omega \geq 0$); *ii)* os pesos devem ter valores bem pequenos pelo que foi adicionado um componente na função de custo que penaliza grandes pesos (regularização L2). Como resultado, a expressão da função $g(x)$ pode ser escrita como

$g(x) = -\delta(x, P_i) = -\omega\,|x - \mu_i| = -\,|\omega x - \omega\mu_i|$ com $\omega \geq 0$. Consequentemente:

$$\frac{\partial g(x)}{\partial \omega} = \begin{cases} \mu_i - x, & \text{se } \omega x - \omega\mu_i \geq 0 \\ x - \mu_i, & \text{se } \omega x - \omega\mu_i < 0. \end{cases} \tag{7.5}$$

Nota-se que μ_i é um vetor constante que representa o vetor das características típicas da i-ésima categoria (protótipo abstrato). Assim, a aprendizagem não é realizada diretamente com o vetor de características da imagem do objeto, senão com a diferença da característica em relação ao protótipo abstrato da i-ésima categoria. Finalmente, se $C(X, y, \omega, b) + \frac{\lambda}{2n}\sum_{\omega}\omega^2$ é a função de custo usada pelo modelo CNN de referência (por exemplo, *negative log-likelihood loss*), os modelos que usam a camada PS-Layer não alteram a função de custo inicialmente usada. Consequentemente, os modelos resultantes de usar a camada PS-Layer podem ser treinados usando as mesmas condições de treinamento que o modelo CNN usado como referência.

Usabilidade da camada PS-Layer em um modelo CNN

Nesta pesquisa assume-se a hipótese de que é possível aumentar a qualidade de classificação dos modelos CNN usando a abordagem de classificação semântica proposta. Assim, para usar a camada *PS-Layer* proposta em um modelo CNN de classificação pré-treinado usado como modelo-CNN-referência, basta seguir os passos definidos no Algoritmo 4 e que podem ser resumidos como:

i) calcular os protótipos semânticos correspondentes ao modelo-CNN-referência usando o Algoritmo 1;

ii) eliminar a camada *softmax* do modelo-CNN-referência;

iii) insertar a camada PS-Layer no final do modelo-CNN-referência;

iv) re-treinar o modelo PS-Layer resultante usando as mesmas condições de treinamento do modelo-CNN-referência.

A Figura 7.1a) apresenta um exemplo de modelo CNN resultante de insertar a camada PS-Layer proposta em um modelo-CNN pré-treinado seguindo os passos apresentados no Algoritmo 4.

Algoritmo 4 Construção de um modelo CNN com a camada PS-Layer

Entrada: Modelo CNN pré-treinado Λ, banco de imagens de objetos O
Saída: Novo Modelo CNN Λ'
$prototypes\_dataset \leftarrow \{\}$
for $c_i \in C$ **do**
    $P_i \leftarrow prototype\_construction(\Lambda, O, c_i)$
    $prototypes\_dataset \leftarrow prototypes\_dataset \cup P_i$
$new\_model \leftarrow \Lambda.remove\_layer(softmax)$
$ps\_layer \leftarrow make\_PS\_Layer(prototypes\_dataset)$
$\Lambda' \leftarrow new\_model.add\_layer(ps\_layer)$
$\Lambda'.train\_model()$
return Λ'

7.4 Experimentos e Resultados

Os experimentos apresentados nesta seção foram desenvolvidos com o intuito de validar o desempenho da camada PS-Layer proposta quando é usada em um modelo-CNN pré-treinado. No processo de avaliação foi analisado o desempenho da qualidade de classificação dos modelos resultantes de usar a PS-Layer em relação com o modelo-CNN-referência usado como ponto de partida. Os experimentos foram realizados usando, como modelos-CNN-referências, cinco modelos CNN de classificação pré-treinados e três bancos de dados de imagens.

7.4.1 Configuração Experimental

Bancos de dados

Nos experimentos realizados foram usados os bancos de dados MNIST (Lecun et al., 1998), CIFAR-10 (Krizhevsky & Hinton, 2009) e CIFAR-100 (Krizhevsky & Hinton, 2009) para a construção dos protótipos semânticos e o processo de treinamento dos novos modelos que usam a camada PS-Layer.

Modelos

Os experimentos nesta seção visam avaliar o desempenho da camada PS-Layer proposta em diferentes cenários. Nota-se que a semântica capturada pelo modelo CPM proposto depende fortemente da qualidade dos protótipos semânticos calculados usando o Algoritmo 1 em um modelo CNN pré-treinado. Observa-se que os protótipos semânticos são construídos usando apenas aquelas imagens identificadas pelo modelo-CNN-referência como *típicas* (provabilidade Top1 $>$ 0.99); assim, os protótipos não são construídos

usando todas as imagens do banco de dados para uma categoria específica e, consequentemente, essa amostra de imagens usadas para calculá-los dependerá da acurácia do modelo-CNN-referência usado. Ou seja, por definição, o protótipo é construído usando uma porcentagem de imagens do banco de dados (*train*) que é proporcional ao valor Top1-acurácia do modelo CNN usado para calcular os protótipos.

Sob essas condições, analisa-se quanto influencia a qualidade dos protótipos construídos (ainda com modelos de baixa acurácia), no desempenho da abordagem de classificação semântica proposta. Com esse fim, os experimentos foram realizados com modelos CNN pré-treinados que possuem: diferente profundidade na arquitetura do modelo, diferente acurácia de classificação, variedade no número de categorias e diferentes características no bando de dados usados para o treinamento. Nos experimentos realizados usaram-se cinco (5) modelos CNN de classificação pré-treinados como *modelos-CNN-referência*, e que utilizam a camada *softmax* sobre o *valor semântico* na tomada de decisão:

a) três (3) modelos CNN simples (*simples-MNIST*, *simples-CIFAR10* e *simples-CIFAR100*) pouco profundos e que foram construídos e treinados nos bancos de imagens com os respectivos nomes;

b) dois (2) modelos CNN profundos (VGG-CIFAR10, VGG-CIFAR100) baseados na arquitetura VGG-16 (Simonyan & Zisserman, 2014) com a adaptação da abordagem de Liu & Deng (2015) para a classificação nos conjuntos de dados CIFAR-10 e CIFAR-100, respectivamente.

Em cada um dos modelos CNN anteriores, foi substituída a camada *softmax* pela camada *PS-Layer* proposta. Os modelos PS-Layer resultantes foram treinados usando as mesmas condições de treinamento (por exemplo *batch size*, *epochs*, *data augmentation*, etc.) do modelo-CNN-referência correspondente. Várias versões do modelo PS-Layer foram utilizadas visando avaliar o desempenho da camada *PS-Layer* proposta. As versões do modelo PS-Layer diferenciam-se no método de inicialização dos pesos do modelo e na função de distância semântica usada na camada PS-Layer.

Foram avaliadas três (3) versões de modelos que usam a *distância prototípica* na camada PS-Layer:

i) *pttype-scratch*: o novo modelo PS-Layer é treinado desde zero usando o mesmo método de inicialização que o modelo-CNN-referência correspondente.

ii) *pttype-freezing*: o novo modelo PS-Layer é treinado a partir da inicialização dos pesos aprendidos no modelo-CNN-referência. No processo de treinamento, todas as camadas anteriores à camada *PS-Layer* não são treinadas (*freezing*);

iii) *pttype-pret-train*: de maneira análoga à versão *pttype-freezing*, mas o novo modelo PS-Layer é treinado completamente.

Também foram avaliadas três (3) versões de modelos PS-Layer que usam a *distância prototípica penalizada* ($\delta_p(o, P_i) = \delta(o, P_i)+$ *penalidade prototípica*) como métrica de distância semântica na camada PS-Layer:

iv) *pttype-scratch-b*: versão estendida do modelo *pttype-scratch* que usa a distância prototípica penalizada;

v) *pttype-freezing-b*: versão estendida do modelo *pttype-freezing* que usa a distância prototípica penalizada;

vi) *pttype-pret-train-b*: versão estendida do modelo *pttype-pret-train* que usa a distância prototípica penalizada.

Os experimentos foram realizados mediante 5 estudos de casos, um para cada uma das cinco (5) arquiteturas de modelos CNN de classificação que utilizam a camada *softmax* na tomada de decisão. Em seguida, foi usada cada uma dessas arquiteturas como modelo-CNN-referência na construção dos modelos CNN que usam a camada PS-Layer (Ver o Algoritmo 4). Em resumo, nos experimentos realizados foi avaliado o desempenho da camada PS-Layer proposta em seis (6) versões diferentes de cada modelo-CNN-referência usado, completando um total de trinta (30) modelos avaliados.

Os hiper-parâmetros $factor$ e $penalty$

As versões do modelo PS-Layer que usam a *distância prototípica penalizada* precisam da correta seleção e ajuste dos hiper-parâmetros constantes *factor* (κ) e penalty (ϕ) que compõem a *penalidade prototípica* (Ver Definição 11). Dependendo da versão do modelo PS-Layer e do banco de dados, foram variados os valores de configuração desses hiper-parâmetros para analisar o possível impacto que pode ter no desempenho de cada novo modelo treinado.

Nos experimentos realizados foram utilizadas diferentes combinações de $factor(\kappa)$ e *penalty* (ϕ). O objetivo desses experimentos é encontrar a combinação dos hiper-parâmetros que alcança o melhor desempenho dos modelos PS-Layer que usam a *distância prototípica penalizada* (versões *iv)*, *v)* e *vi)*) na camada PS-Layer. Cada versão desses modelos foi treinada para todas as combinações dos valores dos hiper-parâmetros $\kappa = [1, \sqrt{2}, 2, 2\sqrt{2}, 3, 3.5]$ e ϕ=$[0.25, 0.5, 1, 1.5, 2, 2.5, 3, 4]$. Foi avaliado o desempenho - en termos de *acurácia*- das 48 instâncias diferentes do modelo. A Tabela 7.1 apresenta a melhor combinação encontrada dos hiper-parâmetros κ e ϕ nos

experimentos realizados para cada versão PS-Layer que usa a *distância prototípica penalizada*.

7.4.2 Avaliação da camada *PS-Layer*

Nesta seção é analisado o comportamento do desempenho da camada *PS-Layer* em cada uns dos estudos de casos construídos. Para cada estudo de caso foram executados 10 treinamentos de cada versão de modelos que usam a camada *PS-Layer* (6 versões). Em cada experimento apresentado expõe-se a arquitetura dos modelos, a profundidade do modelo-CNN-referência, as características do banco de dados, a qualidade dos protótipos usados e as condições de treinamento. Os experimentos apresentados visam avaliar o desempenho de 300 instâncias de modelos que usam a camada PS-Layer, em termos de acurácia média (e desvio padrão), que foram treinados usando diferentes condições de treinamento. Os valores de acurácia média e desvio padrão foram calculados a partir das 10 instâncias treinadas de cada versão do modelo analisada.

Estudo de caso 1: Modelo simples-MNIST

Esse experimento visa avaliar o desempenho da camada proposta em um modelo simples com alta acurácia. Foi construído um modelo pouco profundo, *simples-MNIST*, baseado na arquitetura de LeNet (Lecun et al., 1998) para a classificação de dígitos no banco de dados MNIST (mesmo modelo usado no Capítulo 5). O modelo *simples-MNIST* usado como referência (e todas as versões que usam a camada PS-Layer) foram treinadas sem usar técnicas de aumento de dados (*data augmentation*), com 50 *epochs* e 128 imagens como tamanho de lote (batch-size) (Ver detalhes do modelo no

Modelo	pttype-scratch-b		pttype-freezing-b		pttype-pret-train-b	
	κ	ϕ	κ	ϕ	κ	ϕ
simples-MNIST	3.5	2.5	$\sqrt{2}$	2.5	$2\sqrt{2}$	2.5
simples-CIFAR10	1.0	0.25	1.0	3.0	$2\sqrt{2}$	0.5
simples-CIFAR100	1.0	0.25	1.0	3.0	3.5	0.25
VGG-CIFAR10	3.0	0.25	1.0	3.0	1.0	0.25
VGG-CIFAR100	3.0	0.25	1.0	3.0	$2\sqrt{2}$	0.5

Tabela 7.1: *Os hiper-parâmetros $factor$ e penalty.* A Tabela mostra a configuração dos hiper-parâmetros $factor(\kappa)$ e $penalty(\phi)$ que consegue o melhor desempenho das versões que usam a *distância prototípica penalizada*: *iv) pttype-scratch-b*, *v) pttype-freezing-b* e *vi) pttype-pret-train-b* de cada modelo-CNN-referência listado.

Apêndice B.1). Os protótipos foram construídos com o 95% (Top1-acurácia do modelo simples-MNIST) do banco de dados de treino.

A Tabela 7.2 mostra o desempenho de cada versão do modelo que usa a camada PS-Layer nos bancos de dados de treino (*Train*) e prova (*Test*). O modelo-CNN-referência (simples-MNIST) que usa a camada *softmax* é mostrado na primeira linha da tabela, separado das outras versões listadas. Em cada banco de dados (*Test* e *Train*) é mostrada a acurácia média (*Mean*), o desvio padrão (*Std*) e o máximo valor (*Max*) de acurácia obtido por cada versão de modelo listada.

Percebe-se como as versões do modelo simples-MNIST treinadas desde zero (*pttype-scratch*, *pttype-scratch-b*), ainda sem usar como ponto de inicialização o conhecimento (pesos aprendidos) do modelo-CNN-referência, consegue alcançar um desempenho semelhante ao modelo-CNN-referência. Por exemplo, a versão *pttype-scratch* alcança inclusive o 100% de acurácia no Top5. Nota-se também como todas as versões dos modelos que usam a *distância prototípica penalizada* conseguem melhorar o desempenho das versões que usam somente a *distância prototípica*.

Outro comportamento relevante reside em que nenhuma versão *freezing* (que usa as mesmas característica do modelo original) consegue ganhar em desempenho ao modelo simples-MNIST. Isso é consequência de que essas versões do modelo PS-Layer (*pttype-freezing*, *pttype-freezing-b*) não aprendem novamente a característica do objeto para posicioná-la semanticamente em torno ao protótipo da categoria. Consequentemente, a camada PS-Layer não atualiza a estrutura das características e tenta aprender somente a importância de cada característica unitária na categoria. Ou seja, como as características do modelo não podem ser atualizadas, essas versões do modelo (*freezing*) não alcançam um bom desempenho.

Model	Test				Train	
	Top1		Top5		Top1	Top5
	Mean±Std	Max	Mean±Std	Max	Mean±Std	Mean±Std
simples-MNIST	99.22±.04	99.28	99.99±.005	100	99.83±.02	99.99±.001
pttype-scratch	99.16±.06	99.28	100	100	99.76±.04	99.99±.001
pttype-scratch-b	99.18±.06	99.25	99.99±.003	100	99.82±.03	99.99±.001
pttype-freezing	98.89±.02	98.91	100	100	99.42±.01	99.99±.001
pttype-freezing-b	99.05±.03	99.08	100	100	99.53±.01	99.99±.001
pttype-pre-train	99.20±.03	99.24	99.99±.004	100	99.79±.02	99.99±.001
pttype-pre-train-b	**99.23±.05**	**99.34**	**100**	100	**99.83±.02**	99.99±.001

Tabela 7.2: Acurácia alcançada pelas versões do modelo simples-MNIST que usam a camada PS-Layer proposta. Mostra-se em negrito os valores do melhor desempenho alcançado no banco de dados MNIST.

Model	Test				Train	
	Top1		Top5		Top1	Top5
	Mean±Std	Max	Mean±Std	Max	Mean±Std	Mean±Std
simples-CIFAR10	69.53±2.18	72.11	97.43±.42	98.09	74.32±2.74	98.23±.41
pttype-scratch	73.42±.47	74.05	98.0±.26	98.23	79.57±.76	98.95±.14
pttype-scratch-b	73.71±.71	74.96	98.01±.40	98.75	79.49±.81	98.90±.20
pttype-freezing	64.54±.13	64.80	96.56±.04	96.64	68.53±.09	97.26±.05
pttype-freezing-b	67.44±.08	67.54	97.29±.03	97.33	71.37±.06	98.05±.04
pttype-pre-train	75.84±.62	**76.84**	**98.45±.14**	**98.64**	**82.54±1.06**	**99.23±.11**
pttype-pre-train-b	**75.87±.42**	76.47	98.30±.14	98.55	82.25±.55	99.20±.07

Tabela 7.3: Acurácia alcançada pelas versões do modelo simples-CIFAR10 que usam a camada PS-Layer proposta. Mostra-se em negrito os valores do melhor desempenho alcançado no banco de dados CIFAR10.

Um comportamento contrário ao anterior é observado nas versões *pre-train* do modelo *simples-MNIST*. Essas versões usam como ponto de partida o conhecimento (pesos) do modelo original e têm a liberdade de atualizar a estrutura da característica e seus pesos associados. Os resultados mostram que as abordagens *pre-train* alcançam o melhor desempenho em comparação com todas as versões do modelo *simples-MNIST* que usam a camada PS-Layer. Observa-se também como a versão que usa a *distância prototípica penalizada* consegue um desempenho ainda maior (Top1-acurácia média (99.23%) e Top1-acurácia máxima (99.34%) no banco de dados Test) que o modelo *simples-MNIST* referência. Os resultados mostram que ainda em um modelo simples de alta acurácia, a camada PS-Layer consegue classificar o objeto baseado na similaridade com os protótipos com melhor acurácia que o modelo-CNN-referência que usa a camada softmax.

Estudo de caso 2: Modelo simples-CIFAR10

Os experimentos nesse estudo de caso visam usar um modelo pouco profundo *simples-CIFAR10*, em um banco de dados de imagens de objetos reais com poucas categorias (CIFAR10). O modelo-CNN-referência está baseado na arquitetura *Deep Belief Network* (Krizhevsky & Hinton, 2010), possui poucas camadas e foi treinado sem técnicas de aumento de dados, usando 50 *epochs* e com 32 imagens como tamanho de lote (Ver detalhes do modelo no Apêndice B.2). Os protótipos das 10 categorias foram construídos com o 64% (aproximadamente o Top1-acurácia do modelo simples-CIFAR10) das imagens do banco de dados CIFAR10 rotuladas para treino (Ver os protótipos construídos no Apêndice C.2).

A Tabela 7.3 mostra os resultados alcançados em cada modelo avaliado. Observa-

se, novamente, que as versões do modelo *freezing* -como não atualizam as características do objeto- não possuem um bom desempenho comparado com as outras versões avaliadas. Os resultados mais relevantes nesse experimento, residem no fato de que as 4 versões que atualizam as características do objeto (*pttype-scratch, pttype-scratch-b, pttype-pre-train, pttype-pre-train-b*) conseguem ganhar em desempenho ao modelo de referência. Ou seja, a camada *PS-Layer* nessas versões, enquanto consegue reposicionar os objetos da categoria para alcançar uma organização prototípica, ganha em desempenho (acurácia) à abordagem *softmax*.

De maneira similar aos resultados alcançados nos experimentos do estudo de caso simples-MNIST, todas as versões que usam a *distância prototípica penalizada* ganham em desempenho às versões que usam a *distância prototípica*. Ainda quando as versões treinadas desde zero (*from-scratch*) conseguem ganhar em desempenho ao modelo-CNN-referência (exemplo a versão *pttype-scratch-b* consegue uma diferença de acurácia de +4.18% Top1-Test e +5.17% Top1-Train); o melhor resultado foi obtido pelas versões que usam como ponto de inicialização do modelo os pesos aprendidos pelo modelo-CNN-referência. As versões *pttype-pre-train* e *pttype-pre-train-b* alcançaram o melhor desempenho dentre os modelos treinados no banco de dados CIFAR10, conseguindo aumentar a acurácia nos bancos de dados de prova (+6.34% Top1-Test e +1.02% Top5-Test) e de treino (+8.22% Top1-Train e +1% Top5-Train) com relação ao modelo simples-CIFAR10 (Ver valores em negrito na Tabela 7.3).

Os resultados mostram que as versões *pre-train* possuem uma melhor exatidão na acurácia obtida nas 10 rodadas de treino realizadas para cada versão listada. Por exemplo, observa-se na Tabela 7.3 como o modelo de referência possui uma variação da acurácia (Std) no banco de dados de prova de ±2.18% Top1-Test, enquanto a variação das versões *pre-train* não supera o 0.62% Top1-Test com as mesmas imagens de prova.

Os resultados mostram que a camada PS-Layer na arquitetura simples do modelo utilizado (simples-CIFAR10) consegue, com marcada diferença, superar o desempenho da camada *softmax* na mesma arquitetura e com as mesmas condições de treinamento.

Estudo de caso 3: Modelo simples-CIFAR100

Esse experimento tem como propósito avaliar o desempenho da camada PS-Layer em um modelo simples, com baixa acurácia, mas que usa um banco de dados de imagens de objetos reais com maior número de categorias (CIFAR100). O modelo *simples-CIFAR100* possui a mesma arquitetura que o modelo usado no estudo de caso simples-CIFAR10, mas a última camada está modificada de acordo com o número de categorias (100 categorias). O modelo simples-CIFAR100 foi treinado sem usar técnicas de

Model	Test				Train	
	Top1		Top5		Top1	Top5
	Mean±Std	Max	Mean±Std	Max	Mean±Std	Mean±Std
simples-CIFAR100	47.21±.52	47.88	75.91±.48	76.53	83.42±1.58	96.86±.44
pttype-scratch	49.92±.61	50.70	77.55±.54	78.53	86.13±1.06	97.18±.28
pttype-scratch-b	50.00±.08	50.60	77.63±.77	78.53	85.12±.83	97.06±.49
pttype-freezing	41.84±.13	42.10	71.98±.22	72.30	62.87±.24	88.15±.16
pttype-freezing-b	44.33±.83	44.46	74.02±.06	74.15	67.85±.06	90.74±.03
pttype-pre-train	50.19±.47	50.77	78.00±.18	78.24	**87.73±.74**	98.08±.20
pttype-pre-train-b	**50.24±.30**	**50.80**	**78.09±.51**	**78.96**	87.54±.50	**98.26±.40**

Tabela 7.4: Acurácia alcançada pelas versões do modelo simples-CIFAR100 que usam a camada PS-Layer proposta. Mostra-se em negrito os valores do melhor desempenho alcançado no banco de dados CIFAR100.

aumento de dados, usando 150 *epochs* e com 64 imagens como tamanho de lote (Ver detalhes do modelo no Apêndice B.3). Os protótipos usados como conhecimento a priori da camada PS-Layer foram calculados com o 45% (aproximadamente a acurácia do modelo-CNN-referência) das imagens do banco de treino. Nota-se como, nesse caso, os protótipos foram construídos com menos da metade das imagens que conformam o banco de dados.

A Tabela 7.4 mostra os resultados de acurácia alcançados pelas versões construídas do modelo simples-CIFAR100. Observa-se como o comportamento das versões do modelo que usam a camada PS-Layer é semelhante ao comportamento dos estudos de casos analisados anteriormente. As versões *freezing* continuam sem superar o desempenho alcançado pelo modelo-CNN-referência. Nota-se como –novamente– as versões *pre-train* do modelo alcançam a melhor acurácia dentre as versões treinadas, inclusive melhor acurácia que o modelo-CNN-referência no banco de dado de prova (+3.03% Top1-Test e +2.18% Top5-Test) e de treino (+4.31% Top1-Train e +1.4% Top5-Train).

Estudo de caso 4: Modelo VGG-CIFAR10

Esse experimento visa avaliar o desempenho da camada PS-Layer em um modelo profundo (VGG-CIFAR10), com alta acurácia, mas que usa um banco de dados de imagens de objetos reais com poucas categorias (CIFAR10). O modelo-CNN-referência VGG-CIFAR10 foi construído reproduzindo a arquitetura profunda proposta por Liu & Deng (2015), a qual se inspira na abordagem do modelo VGG16 mas, adaptado ao banco de dados CIFAR10. O modelo de referência (VGG-CIFAR10) e as versões que usam a camada PS-Layer foram treinadas usando as mesmas condições: técnicas de aumento

de dados, usando 250 *epochs* e com 128 imagens como tamanho de lote (Ver detalhes do modelo no Apêndice B.4). Os protótipos das 10 categorias do banco de dados CIFAR10 foram construídos com o 85% das imagens de treino.

A Tabela 7.5 resume os valores da acurácia obtidos para cada um dos modelos usados nos experimentos. Observa-se como a profundidade do modelo de referência VGG-CIFAR10 (60 camadas) não influencia em que as versões do modelo que usam a camada PS-Layer possuam um comportamento similar ao observado nos experimentos dos estudos de casos anteriores.

De maneira similar aos resultados alcançados nos experimentos anteriores, todas as versões que usam a *distância prototípica penalizada* ganham em desempenho às versões que usam a *distância prototípica*. As versões treinadas desde zero (*from-scratch*) também ganham em desempenho ao modelo de referência. Novamente, o melhor resultado foi obtido pelas versões que usam como ponto de inicialização do modelo os pesos aprendidos pelo modelo de referência (*pttype-pre-train*, *pttype-pre-train-b*). Essas versões *pre-train* que usam a camada PS-Layer alcançaram o melhor desempenho dentre os modelos treinados no banco de dados CIFAR10, conseguindo aumentar a acurácia nos bancos de dados de prova com relação ao modelo VGG-CIFAR10 (Ver valores em negrito na Tabela 7.5).

Estudo de caso 5: Modelo VGG-CIFAR100

Esse experimento –de maneira similar ao estudo de caso VGG-CIFAR10– tem como propósito avaliar o desempenho da camada PS-Layer em um modelo profundo, com acurácia razoável e que usa um banco de dados de imagens de objetos reais com mais categorias (CIFAR100). O modelo de referência utilizado VGG-CIFAR100 possui a

Model	Test				Train	
	Top1		Top5		Top1	Top5
	Mean±Std	Max	Mean±Std	Max	Mean±Std	Mean±Std
VGG-CIFAR10	93.43±.17	93.72	99.84±.03	99.86	**99.94±.09**	**100**
pttype-scratch	93.48±.25	93.91	99.74±.04	99.80	99.85±.03	99.99±.001
pttype-scratch-b	93.49±.18	93.80	99.73±.02	99.75	99.84±.04	100
pttype-freezing	93.39±.01	93.39	99.53	99.53	99.99	100
pttype-freezing-b	93.40±.01	93.41	99.60	99.61	99.99	100
pttype-pre-train	93.65±.15	**94.01**	99.75±.03	99.80	99.87±02	99.99±.01
pttype-pre-train-b	**93.88±.15**	93.99	**99.84±.03**	**99.90**	99.85±.02	99.99±.01

Tabela 7.5: Acurácia alcançada pelas versões do modelo VGG-CIFAR10 que usam a camada PS-Layer proposta. Mostra-se em negrito os valores do melhor desempenho alcançado no banco de dados CIFAR10.

mesma arquitetura que o modelo usado no estudo de caso VGG-CIFAR10, mas a última camada está modificada para classificar as 100 categorias do bando de dados. Os protótipos foram construídos usando o 60% das imagens do banco de dados de treino.

A Tabela 7.6 resume os valores de acurácia obtidos nas versões do modelo de referência VGG-CIFAR100. Os valores de acurácia alcançados pelas versões PS-Layer treinadas possuem um comportamento similar aos observados nos estudos de casos apresentados anteriormente. Novamente, as versões *pre-train* do modelo PS-Layer alcançam a melhor acurácia dentre as versões treinadas, mas não conseguem ganhar em desempenho (no Top5 do banco de dados de prova) ao modelo referência.

Análise semântica

Nesta seção analisa-se a qualidade de classificação semântica da abordagem baseada em protótipos proposta em comparação com a abordagem clássica dos modelos CNN de classificação. O intuito desse experimento não é somente analisar a acurácia dos modelos, senão analisar se os membros incorretamente classificados de uma categoria possuem alguma relação semântica com aquelas categorias onde foram categorizados por cada modelo.

Para realizar o experimento foram escolhidos dois estudos de casos –dos apresentados anteriormente– que foram treinados para realizar a classificação no banco de dados de imagens CIFAR10. O critério de escolha desse banco de dados reside em que são poucas categorias e as imagens referem-se a objetos. Para análise foram selecionados dois modelos-CNN-referência com diferenças na profundidade da arquitetura e na acurácia da classificação (simples-CIFAR10 e VGG-CIFAR10).

Model	Test				Train	
	Top1		Top5		Top1	Top5
	Mean±Std	Max	Mean±Std	Max	Mean±Std	Mean±Std
VGG-CIFAR100	70.43±.04	70.76	**90.44±.18**	**90.66**	99.20±.03	99.99±.01
pttype-scratch	69.97±.38	70.44	89.30±.21	89.67	97.14±2.34	99.76±.03
pttype-scratch-b	70.78±.33	71.32	90.17±.10	90.32	99.08±.07	99.98±.00
pttype-freezing	70.55±.01	70.57	89.01±.02	89.04	99.66±.00	100
pttype-freezing-b	70.56±.03	70.60	89.15±.03	89.19	99.69	100
pttype-pre-train	71.28±.20	71.61	90.32±.22	90.60	99.30±.04	99.99±.00
pttype-pre-train-b	**71.42±.27**	**71.79**	**90.41±.12**	90.61	**99.37±.04**	99.99±.00

Tabela 7.6: Acurácia alcançada pelas versões do modelo VGG-CIFAR100 que usam a camada PS-Layer proposta. Mostra-se em negrito os valores do melhor desempenho alcançado no banco de dados CIFAR100.

Para a análise semântica realizada foram agrupadas as categorias do banco de dados CIFAR10 em duas macro-categorias semânticas: *0-veículos de transporte* e *1-animais*. A macro-categoria *0-veículos de transporte* está conformada pelas categorias: 0-avião, 1-automóvel, 8-barco e 9-caminhão. A macro-categoria *1-animais* está constituída pelas categorias: 2-ave, 3-gato, 4-cervo, 5-cachorro, 6-sapo e 7-cavalo. Primeiramente, analisou-se a semântica contida nas representações dos protótipos semânticos construídos e agruparam-se os protótipos –pela sua similaridade estrutural– no mesmo número de grupos como macro-categorias semânticas definidas. A Figura 7.3 mostra o agrupamento hierárquico dos protótipos semânticos construídos com o modelo simples-CIFAR10 no banco de dados CIFAR10. Observa-se que as representações das categorias semânticas distribuem o conjunto de dados CIFAR10 alcançando uma organização semântica hierárquica. Por exemplo, duas macro-categorias são facilmente visíveis na Figura 7.3: animais e veículos de transporte. Os protótipos que compõem cada um dos grupos (*clusters*) construídos coincidem com as macro-categorias semânticas definidas (*0-veículos de transporte*, *1-animais*), o que mostra a qualidade de representação semântica dos protótipos semânticos construídos. Nota-se que a macro-categoria *0-veículos de transporte* também é semanticamente interpretada por nossa representação como duas sub-categorias: veículos não terrestres e veículos terrestres.

Seguidamente analisou-se o comportamento da classificação dos objetos para cada categoria do banco de dados em cada estudo de caso selecionado. Foi usada a matriz de confusão de cada modelo para comparar a qualidade da classificação semântica baseada em protótipos proposta com relação à abordagem tradicional. De forma geral, nesse experimento são analisados dois comportamentos diferentes do método de

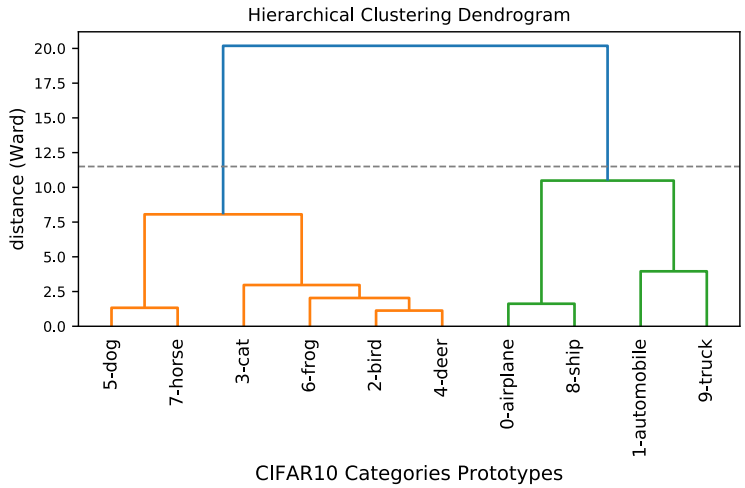


Figura 7.3: Agrupamento hierárquico dos protótipos semânticos construídos com o modelo de referência simples-CIFAR10.

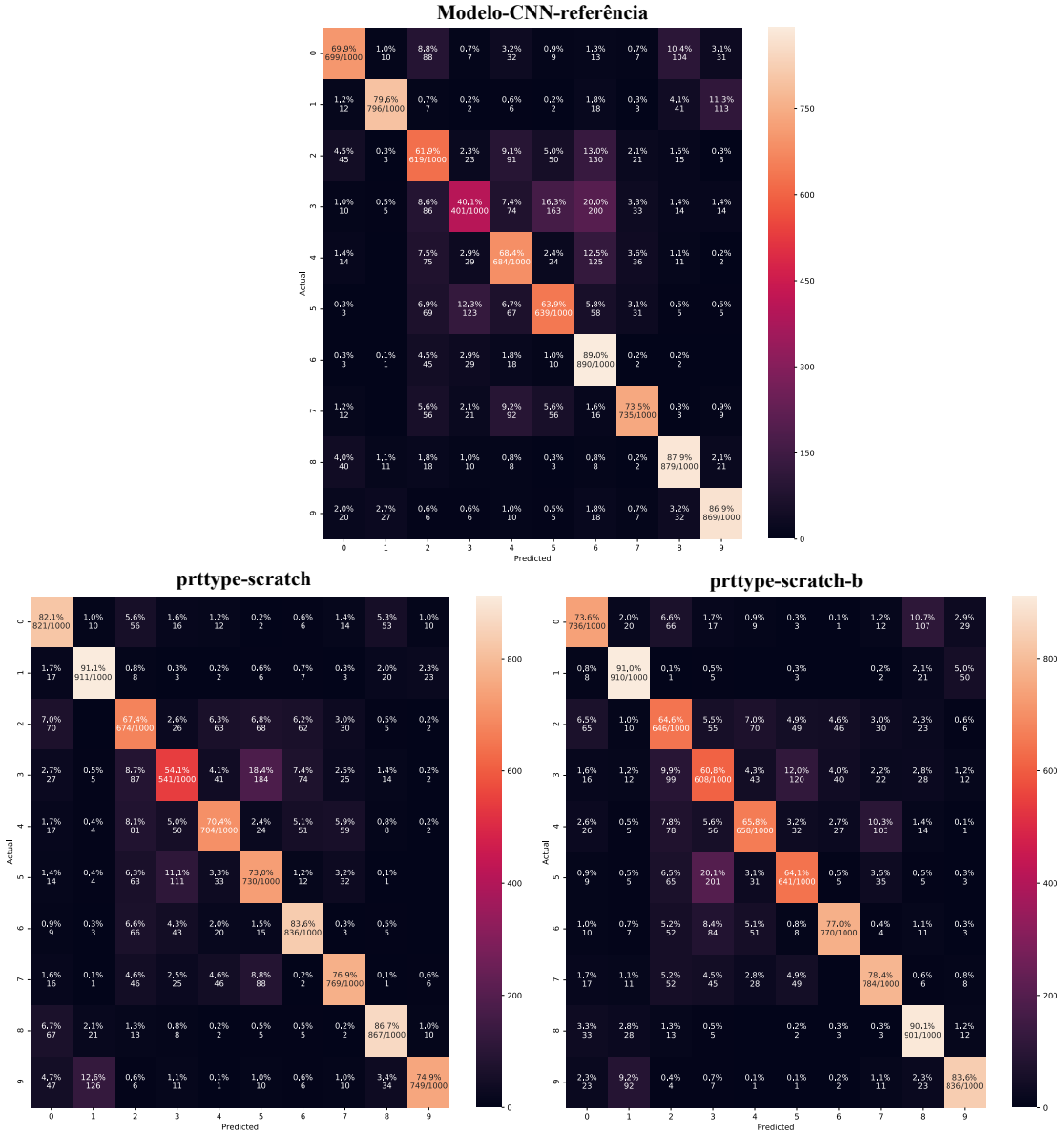


Figura 7.4: *Comparação entre as matrizes de confusão das versões pttype-scratch do modelo simples-CIFAR10 com relação ao modelo-CNN-referência.*

classificação baseado em protótipos: *i)* a capacidade de generalização semântica do método de aprendizagem, usando apenas o protótipo da categoria como conhecimento a priori (versões *pttype-scratch*); *ii)* a capacidade do método de aprendizagem de reajustar a predição semântica do modelo inicial, usando como conhecimento a priori os protótipos semânticos e os pesos aprendidos pelo modelo-CNN-referência (versões *pttype-pre-train*).

A Figuras 7.4 e 7.5 apresentam as matrizes de confusão das versões PS-Layer *pttype-scratch, pttype-scratch-b* e *pttype-pre-train, pttype-pre-train-b* do modelo-CNN-

referência simples-CIFAR10, respectivamente. Observa-se que nesse modelo de baixa acurácia, os protótipos semânticos são construídos usando menos da metade das imagens do banco de dados para cada categoria.

A Figura 7.4 mostra as matrizes de confusão da família de modelos que foram treinadas desde zero *pttype-scratch*. Analisou-se a nova restruturação dos elementos classificados incorretamente em cada categoria. Por exemplo, na categoria 9-caminhão, o novo modelo simples-CIFAR10 *pttype-scratch* perde em acurácia respeito ao modelo-CNN-referência em um 12%, mas esses novos elementos classificados incorretamente foram classificados corretamente dentro da macro-categoria semântica *veículos de transporte*. Inclusive, com menor acurácia, foram classificados menos elementos da categoria 9-caminhão na macro-categoria *animais* (-8 elementos), e observa-se que a maioria (12%) dos elementos classificados incorretamente foram categorizados na categoria 1-automóvel, a categoria mais próxima semanticamente da categoria 9-caminhão.

Aliás, a Figura 7.4 também mostra a versão *pttype-scratch-b* que possui um método de treinamento que penaliza a distância semântica do objeto quando é marcadamente diferente do protótipo da categoria de interesse. Nota-se como nas categorias que maior acurácia possuem (1-automóvel e 8-barco) são realocados os elementos classificados incorretamente da macro-categoria semântica *animais* para a macro-categoria *veículos de transporte*.

A Figura 7.5 mostra as matrizes de confusão da família de versões *pttype-pre-train*, as quais foram treinadas usando como ponto de partida os pesos aprendidos pelo modelo-CNN-referência simples-CIFAR10. Observa-se na categoria 1-automóvel que a versão do modelo *pttype-pre-train* conseguiu aumentar a acurácia e diminuiu em 10 os elementos incorretamente classificados na macro-categoria *animais*. Nesse exemplo, as instâncias da categoria 1-automóvel não foram classificadas como 4-cervo e 7-cavalo.

No caso da categoria 3-gato nota-se que, além de aumentar a acurácia respeito ao modelo-CNN-referência, a versão *pttype-pre-train* aumentou a quantidade de elementos classificados na macro-categoria semântica *animais*. Nesse exemplo percebe-se como, dentro da macro-categoria *aminais*, a percentagem de elementos classificados incorretamente em cada categoria diminuiu, destacando a categoria 6-sapo que diminuiu significativamente en um 16%. Outro dado interessante é que a única categoria que aumentou a percentagem dos elementos incorretamente classificados foi a categoria 5-cachorro, a qual é a mais próxima semanticamente à categoria 3-gato de todas as categorias do banco de dados CIFAR10.

Aliás, avaliou-se a qualidade da classificação semântica de cada um das versões do modelo *simples-CIFAR10*, calculando cada uma das métricas de classificação de cada modelo; mas, tendo em conta a habilidade dos modelos de classificar objetos cor-

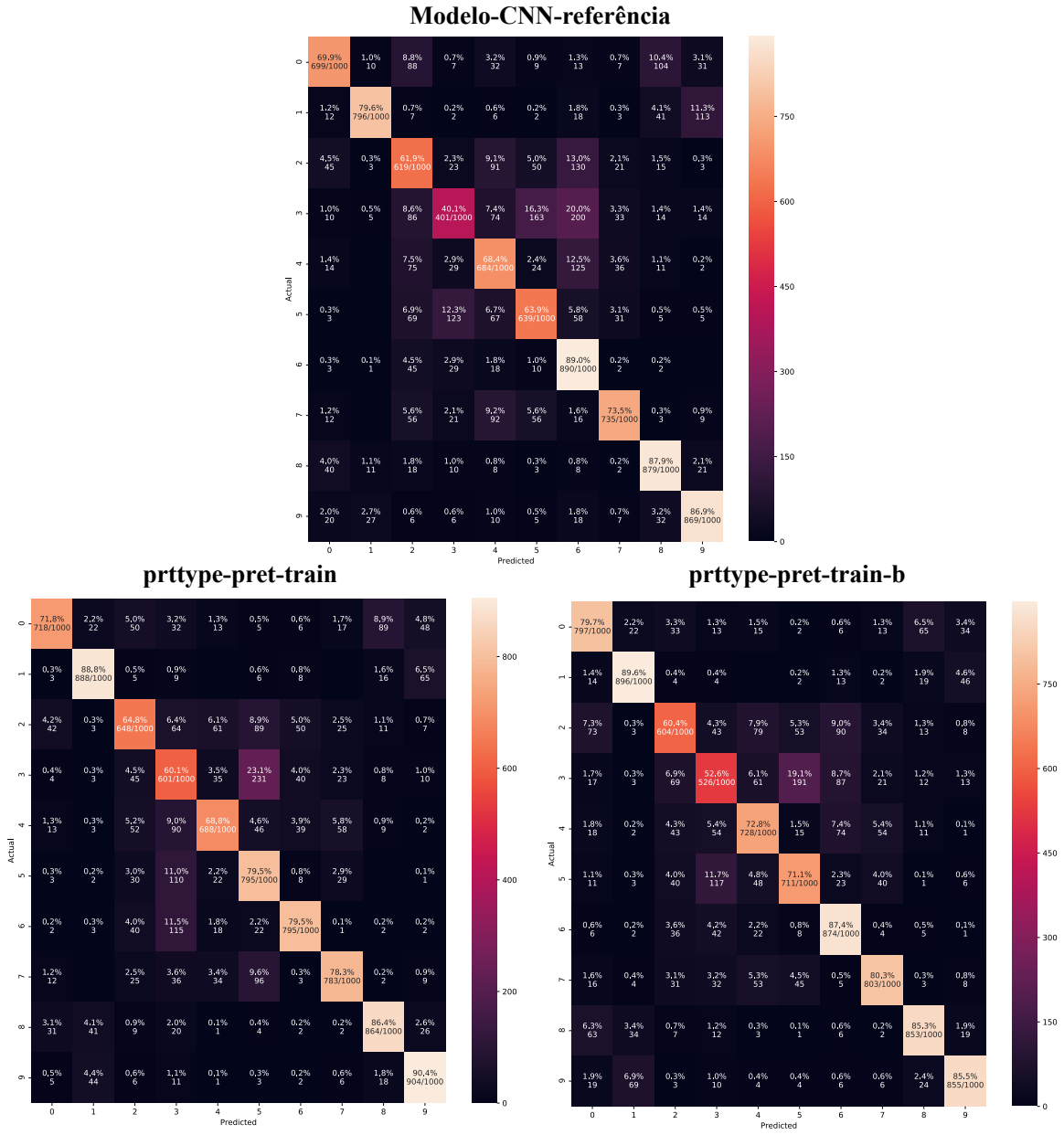


Figura 7.5: *Comparação entre as matrizes de confusão das versões pttype-pre-train do modelo simples-CIFAR10 com relação ao modelo-CNN-referência.*

retamente dentro de cada uma das macro-categorias semânticas construídas. O termo classificação semântica, nesse experimento, refere-se ao fato de classificar corretamente os objetos dentro das macro-categorias semânticas definidas, enquanto também se alcançam boas métricas de classificação dentro das categorias originais. Ou seja, semanticamente, o error de classificar um gato como um cachorro não é considerado um error semântico grave comparado com o erro de classificar o gato como um avião; pois esse último exemplo sim é considerado um erro grave de interpretação semântica da imagem

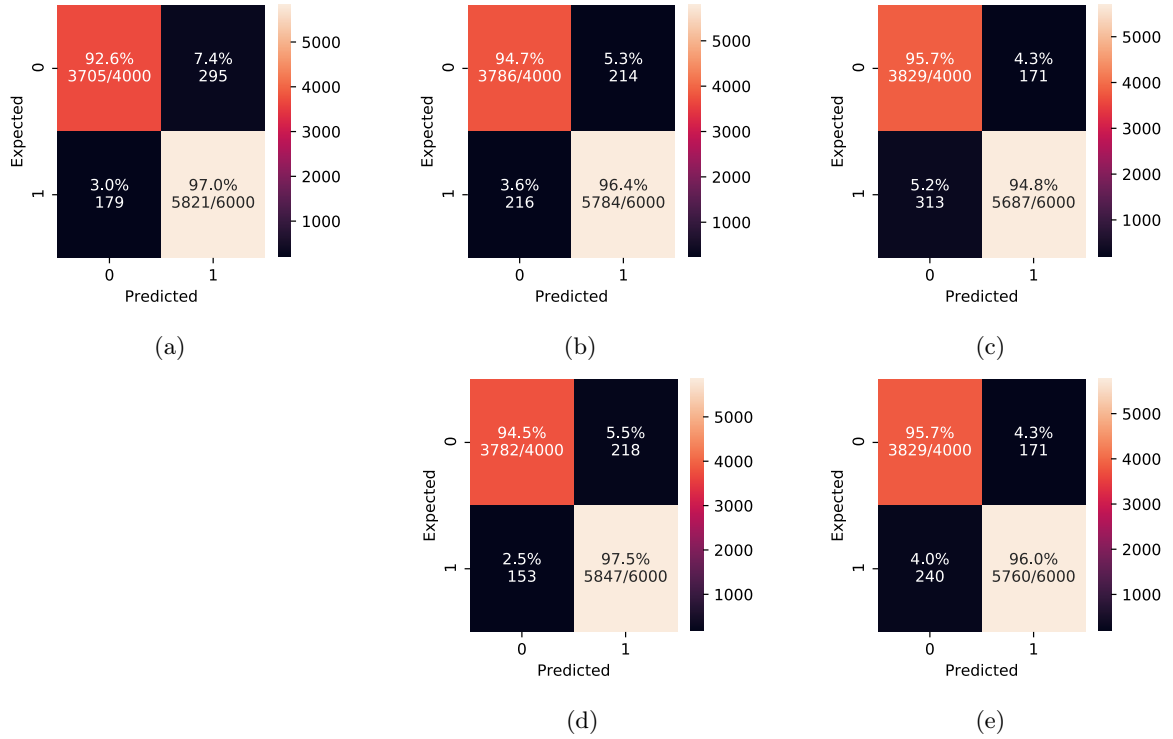


Figura 7.6: Desempenho das versões baseadas em protótipos do modelo simples-CIFAR10 no espaço semântico das macro-categorias *0-veículos de transporte* e *1-animais* do banco de dados CIFAR10. Matrizes de confusão das versões do modelo simples-CIFAR10: *a)* modelo-CNN-referência; *b)* pttype-scratch; *c)* pttype-scratch-b; *d)* pttype-pre-train; e *d)* pttype-pre-train-b.

do gato. A Figura 7.6 apresenta – mediante as matrizes de confusão – o desempenho das versões de classificação baseada em protótipos treinadas desde zero (*pttype-scratch*). Observa-se como as versões *pttype-scratch* do modelo-CNN-referência conseguem, além de melhores métricas de classificação (Ver Tabela 7.3), uma melhor acurácia semântica, pois diminui o erro semântico de classificar *animais* como *veículos de transporte*, e vice-versa. Outro resultado interessante da classificação baseada em protótipos treinadas desde zero (*from scratch*) é que os novos modelos convergem para estados com acurácias próximas ou maiores que a acurácia do modelo original. Ou seja, pode-se disser que, diferente da abordagem de transferir o conhecimento entre modelos CNN usando os pesos aprendidos, o conhecimento do modelo-CNN-referência é transferido para as versões *pttype-scratch* usando –estritamente– os protótipos semânticos das categorias de objetos construídos com o conhecimento do modelo original.

A Figura 7.6 também apresenta as matrizes de confusão das versões *pttype-pre-train* e *pttype-pre-train-b* do modelo simples-CIFAR. Observa-se como ambos os modelos conseguem superar em acurácia ao modelo simples-CIFAR-referência, mas a versão

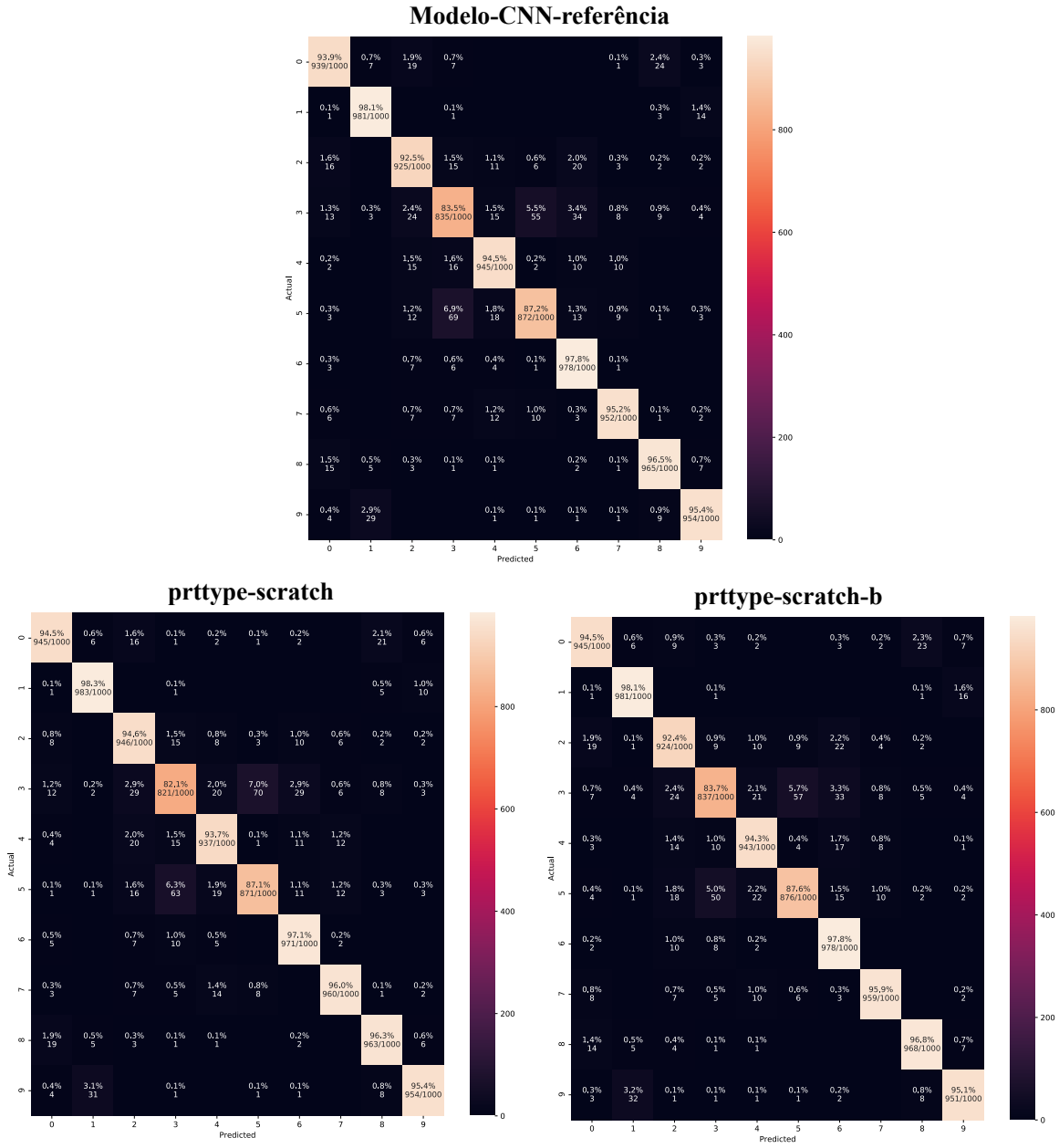


Figura 7.7: *Comparação entre as matrizes de confusão das versões pttype-scratch do modelo VGG-CIFAR10 com relação ao modelo-CNN-referência.*

pttype-pre-train também alcançou o melhor desempenho na classificação das macro-categorias semânticas construídas.

A Figura 7.7 mostra as matrizes de confusão da família *pttype-scratch* correspondentes ao modelo-CNN-referência de alta acurácia e profundidade VGG16-CIFAR10. Esse modelo-CNN-referência tende a confundir a categoria 1-avião com a categoria 2-ave. Observa-se como, além de aumentar a acurácia, as versões *pttype-scratch* e *pttype-scratch-b*, conseguem diminuir a quantidade de elementos incorretamente classificados

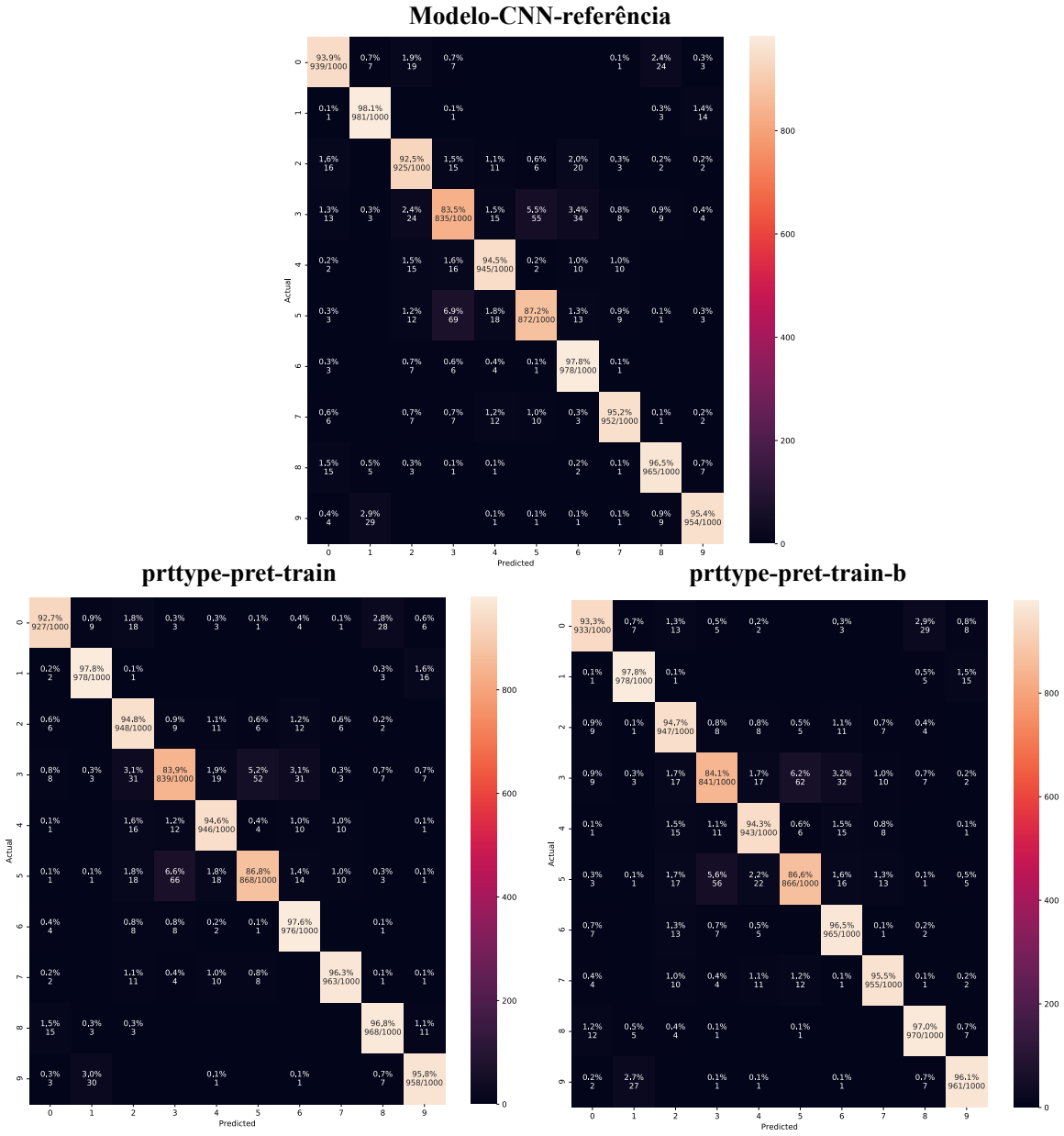


Figura 7.8: *Comparação entre as matrizes de confusão das versões pttype-pre-train do modelo VGG-CIFAR10 com relação ao modelo-CNN-referência.*

na categoria 2-ave. No caso da categoria 3-gato, as versões *pttype-scratch* e *pttype-scratch-b* perdem em acurácia respeito ao modelo-CNN-referência, mas conseguem diminuir os elementos categorizados incorretamente na macro-categoria *0-veículos de transporte* e os consegue realocar na macro-categoria *1-animais*.

Aliás, a Figura 7.8 mostra as versões da família *pttype-pre-train* correspondentes ao modelo-CNN-referência VGG16-CIFAR10. Observa-se como a grande profundidade do modelo-CNN-referência impede que sejam atualizados todos os parâmetros do modelo, mas se podem perceber comportamentos semelhantes aos exemplos ex-

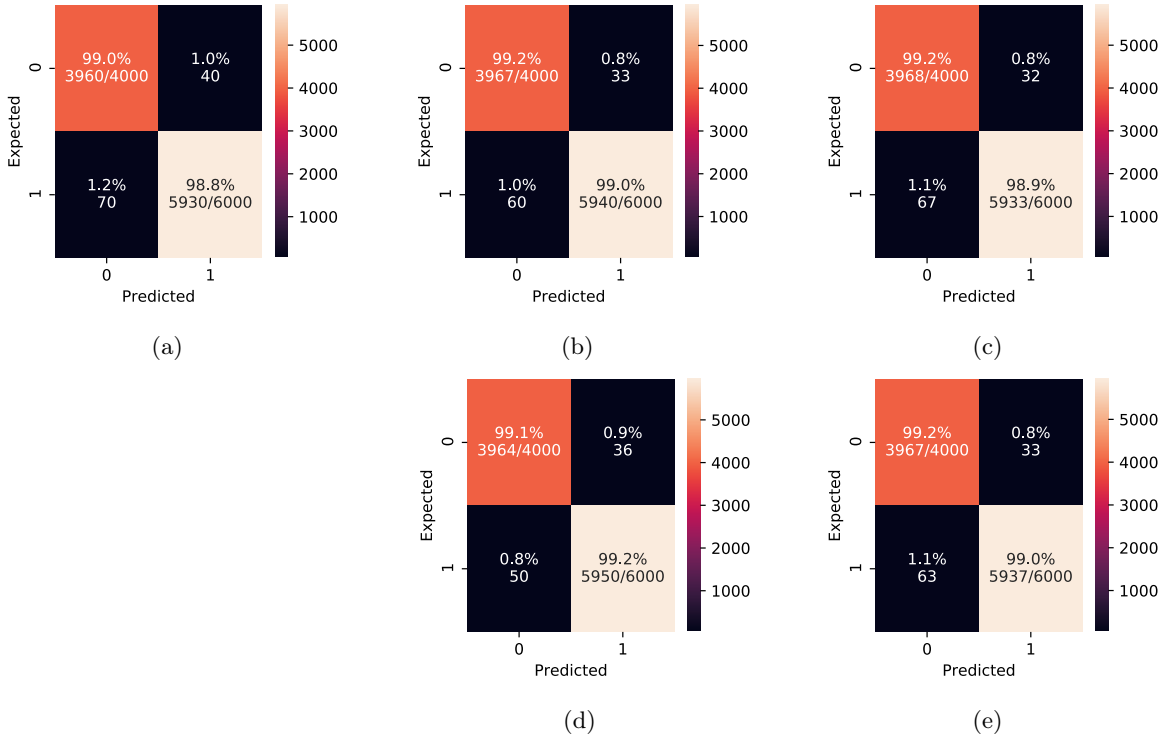


Figura 7.9: Desempenho das versões baseadas em protótipos do modelo VGG-CIFAR10 no espaço semântico das macro-categorias *0-veículos de transporte* e *1-animais* do banco de dados CIFAR10. Matrizes de confusão das versões do modelo VGG-CIFAR10: *a)* modelo-CNN-referência; *b)* pttype-scratch; *c)* pttype-scratch-b; *d)* pttype-pre-train; e *d)* pttype-pre-train-b.

plicados anteriormente. Por exemplo, a categoria 8-barco e a categoria 9-caminhão conseguem ganhar em acurácia ao modelo-CNN-referência e, praticamente, alocam todos os elementos classificados incorretamente na macro-categoria semântica *0-veículos de transporte*.

Analogamente ao estudo de caso do modelo *simples-CIFAR10*, analisou-se o comportamento da qualidade da classificação semântica de cada um das versões do modelo *VGG-CIFAR10*. A Figura 7.9 apresenta as matrizes de confusão –no domínio das macro-categorias semânticas construídas– das versões de classificação baseadas em protótipos treinadas desde zero (*pttype-scratch*); e as versões dos modelos re-treinados a partir dos pesos aprendidos no modelo VGG-CIFAR10-referência (*pttype-pre-train*). Observa-se como, mesmo com um modelo altamente acurado e de arquitetura profunda como o modelo *VGG-CIFAR10*, as versões *pttype-scratch* conseguem convergir para estados com um desempenho semelhante ou melhor que o modelo-CNN-referência. Nota-se que a versão *pttype-pre-train* alcança, novamente, o melhor desempenho dentre todas as versões que usam a camada PS-Layer.

De maneira geral, a abordagem de classificação baseada em protótipos proposta, além de conseguir igual/melhor desempenho em termos de acurácia de classificação, consegue que aqueles elementos incorretamente classificados sejam atribuídos a categorias mais próximas semanticamente com a categoria à qual pertencem. Precisa-se realizar outros tipos de experimentos para analisar com maior profundidade a semântica representada pelo método de classificação semântica proposto neste capítulo.

Avaliação geral da camada PS-Layer

Nos estudos de casos analisados foram avaliadas 6 versões de modelos que usam a camada PS-Layer para a tomada de decisão baseada na similaridade prototípica. Cada uma das versões apresentadas são modificações da arquitetura de um modelo CNN de classificação (que usa a softmax) usado como modelo-CNN-referência. O critério de construção das versões do modelo de referência visa avaliar a camada PS-Layer em modelos com diferente profundidade, diferente acurácia, mudando a inicialização de treinamento realizado em cada novo modelo PS-Layer, e mudando o tipo de distância semântica (*distância prototípica, distância prototípica penalizada*) usada na camada PS-Layer. As versões dos modelos PS-Layer foram treinadas de três formas diferentes: *i)* desde zero (*pttype-scratch, pttype-scratch-b*), *ii)* usando os pesos aprendidos do modelo-CNN-referência sem treinar as camadas anteriores à camada PS-Layer (*pttype-freezing, pttype-freezing-b*) e *iii)* com o treinamento de toda a arquitetura (*pttype-pre-train, pttype-pre-train-b*).

A Figura 7.10 resume o desempenho, em termos de acurácia, de cada uma das versões construídas (*scratch, scratch-b, freezing, freezing-b, pre-train, pre-train-b*) com relação ao modelo-CNN-referência (*baseline*) usado em cada estudo de caso analisado. Na Figura 7.10 cada círculo mostrado corresponde a uma das métricas de acurácia (Test-Top1, Test-Top5, Train-Top1, Train-Top5) apresentadas em cada estudo de caso. O vértice de cada polígono dentro do círculo representa os valores de acurácia (normalizados entre os valores [0-1]) de cada versão no estudo de caso analisado (os nomes do estudo de caso são apresentados em diminutivo). O Apêndice K mostra os exemplos do histórico de treinamento das instâncias dos modelos-CNN-referência e das versões PS-Layer correspondentes.

Os resultados apresentados na Figura 7.10 mostram que as versões que não aprendem a característica do objeto (*freezing, freezing-b*) não conseguem superar o desempenho do modelo-CNN-referência (em preto). Aliás, as quatro versões PS-Layers que aprendem a característica do objeto (famílias *from-scrach* e *pre-train*) - de maneira geral- conseguem um melhor desempenho que o modelo-CNN-referência. As

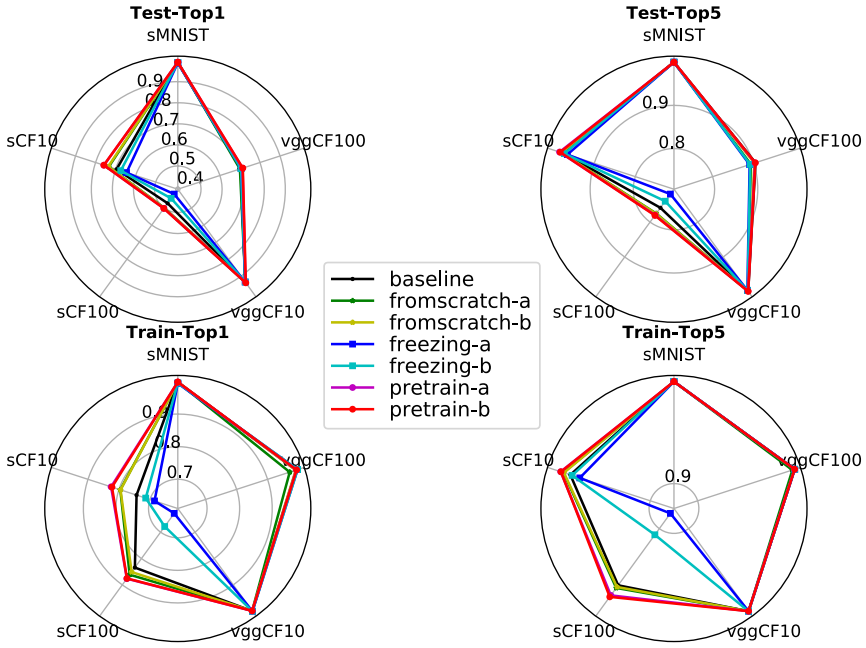


Figura 7.10: *Resumo geral do desempenho da camada PS-Layer.* Apresenta-se o resumo do desempenho (em termos de acurácia) das versões do modelo que usam a camada PS-Layer. Cada círculo resume as métricas de acurácia (Test-Top1, Test-Top5, Train-Top1, Train-Top5) alcançadas em cada um dos estudos de casos analisados. Os valores da acurácia foram normalizados entre os valores [0-1].

versões *pre-train* (em magenta e vermelho) sempre conseguem incrementar a acurácia do modelo-CNN-referência em cada estudo de caso e em cada banco de dados usado nos experimentos. De fato, existem alguns casos de estudo onde a versão *pre-train-b* consegue um incremento significativo ($>$+6%) da acurácia com relação ao modelo-CNN-referência.

As versões do modelo PS-Layer que usam a *distância prototípica penalizada*, geralmente conseguem um melhor desempenho que aquelas versões PS-Layer baseadas na *distância prototípica*. Nota-se que os resultados apresentados nessas versões que usam a distância penalizada foram alcançados com as configurações dos hiper-parâmetros ($factor(\kappa)$ e $penalty(\phi)$) listados na Tabela 7.1. Esses hiper-parâmetros usados constituem a configuração que alcançou melhor desempenho dentre as 42 configurações diferentes usadas; mas isso não significa que sejam as melhores configurações dos hiper-parâmetros para esses modelos que usam a distância prototípica penalizada.

A Figura 7.10 mostra que nos modelos-CNN-referência pouco profundos e de baixa acurácia usados (*simples-CIFAR10* e *simples-CIFAR100*) é onde as versões PS-Layer conseguem um maior incremento da acurácia. Isso constitui um resultado interessante, pois precisamente nesses modelos (de baixa acurácia) os protótipos são construídos com menos imagens do banco de dados de treino. Uma hipótese para

justificar esse comportamento reside em que quando a acurácia é baixa existe maior quantidade de elementos não categorizados corretamente. Consequentemente, existe maior probabilidade de que aumente a acurácia da categorização baseada nas características típicas encapsuladas no protótipo. Essa hipótese é descartada se é analisado o caso do modelo VGG-CIFAR100 que possui baixa acurácia, mas o modelo possui uma arquitetura profunda. Esses pressupostos deixam aberto o debate para analisar a influência da profundidade do modelo de referência selecionado no desempenho das versões que usam a PS-Layer.

Os resultados experimentais mostram que a camada PS-Layer proposta para a tomada de decisão baseada na similaridade prototípica do objeto, não deteriora o desempenho de classificação de um modelo-CNN-referência determinado. A camada PS-Layer introduz o modelo CPM proposto e os conceitos principais da Teoria dos Protótipos nas arquiteturas CNN. Assim, a camada PS-Layer além de aumentar o poder interpretativo desses métodos de aprendizagem profundo, pode também incrementar a acurácia de um modelo-CNN-referência quando é usada como camada de decisão da rede CNN.

7.5 Discussão

No presente capítulo foi proposta uma abordagem de classificação de imagens de objetos que categoriza a imagem de entrada na categoria do protótipo semântico com mais similaridade com a imagem de entrada. Nota-se que a metodologia proposta neste capítulo pode ser entendida como uma tarefa de "recuperação do protótipo mais representativo da imagem de entrada", e constitui um processo da metodologia geral proposta para descrever semanticamente as características globais dos objetos (Ver Figura 4.1). A abordagem apresentada pode ser entendida como um método de *classificação semântica* que classifica o objeto baseado na comparação de todas as características unitárias que o compõem com relação a todas as características representativas (típicas) da categoria encapsuladas no protótipo semântico.

Os resultados obtidos com o método de classificação baseado em protótipos proposto nos estudos de casos realizados, motivam pensar que possa melhorar o desempenho de outros modelos CNN de classificação mais complexos em outros bancos de dados de imagens. Nessa instância da avaliação do método proposto, os resultados apresentados neste capítulo permitem concluir parcialmente que:

i) dentre todas as versões dos modelos PS-Layer, as versões *pre-train* alcançaram o melhor desempenho com relação a todos os modelos avaliados (inclusive o modelo-

CNN-referência). Mesmo quando as versões que realizam o treinamento desde zero (*from-scratch*) conseguem bom resultados, realizar a inicialização com os pesos aprendidos do modelo-CNN-referência, além de alcançar um melhor desempenho, permite usar a camada PS-Layer como último ajuste (*fine tuning*) do modelo inicial;

ii) o desempenho da camada *PS-Layer* proposta justifica que possa ser usada como último passo no processo de treinamento de um modelo CNN de classificação. Ou seja, após de ter um modelo treinado usando a função *softmax*, é suficiente seguir os passos propostos no Algoritmo 4 para usar a camada *PS-Layer* proposta e obter um modelo final com um desempenho melhor que o modelo-CNN-referência;

iii) a aprendizagem baseada na tipicidade do objeto, através da *organização prototípica* da categoria, consegue aumentar o poder interpretativo da tomada de decisões dos modelos CNN de classificação; e também pode melhorar o desempenho de um modelo CNN de classificação que usa a função *softmax* na tomada de decisões.

O método de classificação proposto pode ser facilmente adaptável a qualquer arquitetura dos modelos CNN de classificação, pois somente modifica como deve ser interpretada a semântica do objeto na última camada do modelo CNN. Assim, a camada PS-Layer proposta introduz o *conceito de categorização baseado em protótipos* desenvolvido por Rosch (Rosch, 1975b; Rosch & Mervis, 1975) e propõe uma alternativa de como introduzir a *Teoria dos Protótipos* na tomada de decisões das Redes Neuronais Convolucionais.

Capítulo 8

Conclusões

A *Semântica* como ciência surgiu pela necessidade de analisar a representação e interpretação dos significados atribuídos às palavras, à estrutura das sentenças, aos sinais, aos signos, aos objetos. Em consequência, quando se alude à análise semântica, característica semântica ou representação semântica, refere-se à explicação dos significados atribuídos a um determinado fenômeno e como pode ser interpretado. Ou seja, os significados não são construídos de maneira aleatória e é necessário entender as leis que o governam.

Motivados por como os seres humanos representam e relacionam os significados atribuídos aos objetos, esta pesquisa baseou-se na Teoria dos Protótipos para propor um modelo de representação semática de imagens de objetos. Foi proposto o Modelo Computacional do Protótipo (CPM) que visa representar a estrutura semântica interna das categorias de objetos baseado nos fundamentos da Teoria dos Protótipos. Os experimentos realizados mostraram que o modelo CPM consegue encapsular as características relevantes da categoria no protótipo semântico. Também se mostrou que a métrica de distância semântica proposta consegue simular as ligações semânticas, em termos de tipicidade visual, existentes entre os objetos que compõem a categoria. Ou seja, o modelo CPM proposto pode capturar a tipicidade visual do objeto e o significado central e periférico das categorias de objetos.

Baseado nos resultados obtidos do modelo CPM, foi proposto um modelo de descrição semântica de objetos que usa os componentes essenciais desse modelo (protótipo semântico + métrica de distância semântica) para construir uma representação semântica da imagem do objeto. O Modelo de Descrição baseado em Protótipos proposto usa os protótipos semânticos do modelo CPM para construir uma assinatura discriminativa que descreve semanticamente o objeto destacando as suas características distintivas dentro da categoria.

O Descritor Semântico Global baseado em Protótipos (GSDP) proposto introduz uma nova abordagem de descrição semântica de imagens de objetos. O descritor GSDP não precisa ser treinado e é facilmente adaptável para ser usado com qualquer modelo CNN de classificação. Como foi mostrado nos experimentos realizados no banco de dados ImageNet com os modelos VGG16 e ResNet50, as assinaturas do descritor GSDP são discriminativas, de pequena dimensionalidade e codificam a informação semântica do objeto com relação à categoria à qual pertence. Também foi mostrado que as representações GSDP das imagens de objetos são interpretáveis, pois preservam na sua taxonomia o significado semântico do objeto e a pontuação da tipicidade do objeto dentro da categoria. O Modelo de Descrição baseado em Protótipos apresentado propõe um ponto de partida para introduzir os fundamentos teóricos relacionados à representação do significado semântico da Teoria dos Protótipos na família de Descritores-CNN.

Também foi apresentada uma abordagem de classificação de objetos que usa o método de representação semântica de categorias do modelo CPM proposto para introduzir o conceito de categorização baseado em protótipos nos modelos CNN de classificação. A metodologia proposta usa a Camada de Similaridade Prototípica (PS-Layer) proposta para introduzir todos os conceitos da Teoria dos Protótipos nos modelos CNN classificação: *i)* o protótipo como representação semântica que encapsula as características representativas da categoria; *ii)* a importância relativa das características dos objetos para cada categoria; *iii)* a distância semântica entre os objetos e o protótipo da categoria; e *iv)* a aprendizagem de categorias de objetos baseada na tipicidade visual. Os resultados obtidos com o método de classificação baseado em protótipos proposto mostram que consegue melhorar o desempenho dos modelos CNN escolhidos como estudos de casos, e motivam pensar que é possível melhorar o desempenho de outros modelos CNN de classificação mais complexos usando a abordagem proposta. A camada PS-Layer proposta introduz o conceito de categorização baseado em protótipos e propõe uma alternativa de como introduzir a aprendizagem de conceitos visuais da Teoria dos Protótipos nas Redes Neuronais Convolucionais.

8.1 Limitações da pesquisa

- O modelo CPM proposto não foi validado em bancos de imagens com anotações de tipicidade dos objetos para cada categoria. A impossibilidade de ter dados com essas informações impede realizar uma validação estatística mais robusta do modelo CPM proposto;

- O descritor GSDP proposto foi construído estritamente para descrever imagens

de objetos, e não para descrever cenas. Nota-se que mesmo quando o descritor foi avaliado em imagens que representam cenas, o modelo de representação semântica proposto consegue ganhar em desempenho a outras representações globais da imagem;

- O método de classificação baseado em protótipos proposto requere de protótipos previamente calculados como conhecimento a priori de seus neurônios. Consequentemente, essa abordagem de classificação, ao usar protótipos construídos com outro modelo CNN e que não são atualizados, não pode ser usado como um método de aprendizagem por si só, mas sim como um último passo de treinamento de um modelo CNN de classificação.

8.2 Trabalhos Futuros

- Construir um banco de dados de imagens de objetos com as anotações de tipicidade segundo a interpretação de pessoas, isso permitiria realizar uma avaliação mais robusta do modelo CMP proposto baseado no critério interpretativo dos humanos;

- Avaliar a possibilidade de construir um banco de dados de imagens de objetos com as anotações de tipicidade, por meio da geração de imagens de forma sintética;

- Analisar e comparar a abordagem proposta com outras formas de seleção do protótipo da categoria. Modelar/Realizar experimentos usando mais de um protótipo por categoria. Comparar os resultados da abordagem da Teoria dos Exemplares (Minda & Smith, 2002) com os resultados da Teoria dos Protótipos;

- Propor um método de representação semântica de cenas baseado nas representações (dos objetos que a compõem) construídas com o descritor GSDP proposto. Isso permitiria avaliar o desempenho do descritor GSDP em tarefas de recuperação de imagem (*image retrieval*) e compreensão de cenas (*scene understanding*);

- Avaliar o método de classificação baseado em protótipos proposto em modelos CNN de maior complexidade (VGG16 e ResNet50) e em bancos de dados de imagens de objetos de maior tamanho (ImageNet e Coco);

- Modificar a metodologia de classificação proposta para que o protótipo seja atualizado ciclicamente no processo de treinamento do modelo CNN e o método de classificação baseado em protótipos proposto não precise de métodos de classificação pré-treinados como conhecimento a priori;

- Analisar o uso da metodologia de classificação proposta em ambientes/cenários dinâmicos onde o número de categorias é dinâmico (eliminação/incrementação de categorias no banco de dados). Avaliar o uso da abordagem de protótipos em métodos de classificação como *Zero-Shot Learning*, *Few-Shot learning* e *OpenSet*;

- Analisar o uso da metodologia de classificação baseada em protótipos na tarefa de reconhecimento de faces. Avaliar a posibilidade de construir protótipos de faces usando pontos faciais extraídos com outras bibliotecas (Exemplo Dlib (King, 2009));

- Modificar a arquitetura da camada PS-Layer proposta para que a aprendizagem das categorias de objetos seja realizada enquanto a visualização do protótipo seja também aprendida;

- Analisar o uso da metodologia de classificação baseada em protótipos na aprendizagem hierárquica das categorias de objetos. Avaliar a possibilidade de usar uma abordagem baseada em hierarquia de protótipos.

Referências Bibliográficas

Aanæs, H.; Dahl, A. L. & Pedersen, K. S. (2012). Interesting interest points. *International Journal of Computer Vision (IJCV)*, 97(1):18--35.

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433--459.

Ackley, D. H.; Hinton, G. E. & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147--169.

Adriaens, G. (1993). Process linguistics: a cognitive-scientific approach to natural language understanding. *Conceptualizations and Mental Processing in Language*, 3:141–142.

Agrawal, M.; Konolige, K. & Blas, M. (2008). Censure: Center surround extremas for realtime feature detection and matching. *Proceedings of the of the European Conference on Computer Vision (ECCV)*, pp. 102--115.

Agrawal, P.; Carreira, J. & Malik, J. (2015). Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 37--45.

Alahi, A.; Ortiz, R. & Vandergheynst, P. (2012). Freak: Fast retina keypoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 510--517. Ieee.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175--185.

Ambai, M. & Yoshida, Y. (2011). Card: Compact and real-time descriptors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 97--104. IEEE.

Arthur, D. & Vassilvitskii, S. (2006). How slow is the k-means method? In *Proceedings of the twenty-second annual symposium on Computational geometry*, pp. 144--153. ACM.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, 2:89--195.

Barnes, J. et al. (1995). *The Cambridge Companion to Aristotle*. Cambridge University Press.

Bay, H.; Ess, A.; Tuytelaars, T. & Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding (CVIU)*, 110(3):346--359.

Bendale, A. & Boult, T. E. (2016). Towards open set deep networks. In *CVPR*, pp. 1563--1572.

Berlin, B. & Kay, P. (1991). *Basic color terms: Their universality and evolution (Second Edition)*. Univ. of California Press.

Binder, J. R.; Conant, L. L.; Humphries, C. J.; Fernandino, L.; Simons, S. B.; Aguilar, M. & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130--174.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Berlin.

Boleda, G. & Herbelot, A. (2016). Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, 42(4):619--635.

Breiman, L. (1996). Bias, variance, and arcing classifiers. *Statistics*.

Bristow, H.; Valmadre, J. & Lucey, S. (2015). Dense semantic correspondence where every pixel is a classifier. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4024--4031.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E. & Shah, R. (1994). Signature verification using a"siamese"time delay neural network. In *Advances in Neural Information Processing Systems*, pp. 737--744.

Brown, M.; Hua, G. & Winder, S. (2011). Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(1):43--57.

Bruhn, A.; Weickert, J. & Schnörr, C. (2005). Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision (IJCV)*, 61(3):211--231.

Cadieu, C. F.; Hong, H.; Yamins, D. L.; Pinto, N.; Ardila, D.; Solomon, E. A.; Majaj, N. J. & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963.

Chatfield, K.; Philbin, J. & Zisserman, A. (2009). Efficient retrieval of deformable shape classes using local self-similarities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 264--271. IEEE.

Chebyshev, P. (1867). Des valeurs moyennes. *Journal de Mathematiques pures et Appliquees*, 12:177--184.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K. & Yuille, A. L. (2016). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*.

Chen, X.; Mottaghi, R.; Liu, X.; Fidler, S.; Urtasun, R. & Yuille, A. (2014). Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1971--1978.

Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*.

Chopra, S.; Hadsell, R. & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 539--546. IEEE.

Choy, C. B.; Gwak, J.; Savarese, S. & Chandraker, M. (2016). Universal correspondence network. In *Advances in Neural Information Processing Systems*, pp. 2414--2422.

Cichy, R. M.; Khosla, A.; Pantazis, D. & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346--358.

Collins, J. A. & Curby, K. M. (2013). Conceptual knowledge attenuates viewpoint dependency in visual object recognition. *Visual Cognition*, 21(8):945--960.

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 160--167. ACM.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273--297.

Crammer, K.; Gilad-Bachrach, R.; Navot, A. & Tishby, N. (2003). Margin analysis of the lvq algorithm. In *Advances in Neural Information Processing Systems*, pp. 479--486.

Cuenca, M. J. & Hilferty, J. (1999). *Introducción a la lingüística cognitiva*. Grupo Planeta (GBS).

Dalal, N. & Triggs, B. (2005a). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 886--893. IEEE.

Dalal, N. & Triggs, B. (2005b). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 886--893. IEEE.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E. & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 647--655.

Donoser, M. & Bischof, H. (2006). Efficient maximally stable extremal region (mser) tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 553--560. IEEE.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D. & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2758--2766.

Erhan, D.; Szegedy, C.; Toshev, A. & Anguelov, D. (2014). Scalable object detection using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2147--2154.

Erk, K. (2016). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9:17--1.

Estes, W. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, 115(2):155.

Fei-Fei, L.; Fergus, R. & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(4):594--611.

Fernández, F. & Isasi, P. (2004). Evolutionary design of nearest prototype classifiers. *Journal of Heuristics*, 10(4):431--454.

Fischer, P.; Dosovitskiy, A. & Brox, T. (2014). Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*.

Forsyth, D. & Ponce, J. (2011). *Computer vision: a modern approach*. Upper Saddle River, NJ; London: Prentice Hall.

Fromkin, V.; Rodman, R. & Hyams, N. (2018). *An introduction to language*. Cengage Learning.

Fukushima, K. & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pp. 267--285. Springer.

Fuster, J. M. (1997). Network memory. *Trends in neurosciences*, 20(10):451--459.

Gatys, L.; Ecker, A. & Bethge, M. (2015). A neural algorithm of artistic style. *Nature Communications*.

Geeraerts, D. (1989). Introduction: Prospects and problems of prototype theory. *Linguistics*, 27(4):587--612.

Geeraerts, D. (1993). Des deux côtés de la sémantique structurale: sémantique historique et sémantique cognitive. *Histoire épistémologie langage*, 15(1):111--129.

Geeraerts, D. (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford University Press.

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.

Geiger, A.; Lenz, P.; Stiller, C. & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231--1237.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1440--1448.

Girshick, R.; Donahue, J.; Darrell, T. & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580--587.

Gong, Y.; Lazebnik, S.; Gordo, A. & Perronnin, F. (2013). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2916--2929.

Gong, Y.; Wang, L.; Guo, R. & Lazebnik, S. (2014). Multi-scale orderless pooling of deep convolutional activation features. In *Proceedings of the of the European Conference on Computer Vision (ECCV)*, pp. 392--407. Springer.

Goodman, N. D.; Tenenbaum, J. B. & Gerstenberg, T. (2014). Concepts in a probabilistic language of thought. *To appear in The Conceptual Mind: New Directions in the Study of Concepts*.

Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S. & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187:27--48.

Guzman-Arenas, A. (1968). Computer recognition of three-dimensional objects in a visual scene. Relatório técnico, Massachusetts Institute of Technology (MIT).

Ham, B.; Cho, M.; Schmid, C. & Ponce, J. (2016). Proposal flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3475--3484.

Han, K.; Rezende, R. S.; Ham, B.; Wong, K.-Y. K.; Cho, M.; Schmid, C. & Ponce, J. (2017). Scnet: Learning semantic correspondence. *arXiv preprint arXiv:1705.04043*.

Han, X.; Leung, T.; Jia, Y.; Sukthankar, R. & Berg, A. C. (2015). Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3279--3286.

Haralick, R. M.; Shanmugam, K.; Dinstein, I. et al. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, 3(6):610--621.

He, K.; Zhang, X.; Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770--778.

Hebb, D. (1949). *The organization of behavior; a neuropsychological theory.* Wiley.

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, 93(1):10.

Hinton, G. E.; Osindero, S. & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527--1554.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I. & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735--1780.

Hoijer, H. & Fearing, F. (1954). Language in culture. In *Proceedings of Conference on the Interrelations of Language and Other Aspects of Culture*. Chicago : University of Chicago Press.

Homa, D. & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, 2(3):322.

Horn, B. K. & Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185--203.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M. & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE transactions on information theory*, 8(2):179--187.

Hubel, D. H. & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574--591.

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193--218.

Jaderberg, M.; Simonyan, K.; Zisserman, A. et al. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pp. 2017--2025.

Jaén, J. F. (2006). Breve historia de la semántica histórica. *Interlingüística*, 27(17):345--354.

Jahrer, M.; Grabner, M. & Bischof, H. (2008). Learned local descriptors for recognition and matching. In *Computer Vision Winter Workshop*, volume 2.

Jetley, S.; Romera-Paredes, B.; Jayasumana, S. & Torr, P. (2015). Prototypical priors: From improving classification to zero-shot learning. In *Proceedings of the of the British Machine Vision Conference (BMVC)*.

Jordan, M. (1986). Serial order: a parallel distributed processing approach. technical report, june 1985-march 1986. Relatório técnico, California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science.

Kanazawa, A.; Jacobs, D. W. & Chandraker, M. (2016). Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3253--3261.

Karpathy, A. (2017). Cs231n convolutional neural networks for visual recognition.

Karpathy, A. & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128--3137.

Kaufman, L. & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Khaligh-Razavi, S.-M. & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915.

Kim, J.; Liu, C.; Sha, F. & Grauman, K. (2013). Deformable spatial pyramid matching for fast dense correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2307--2314.

Kim, S.; Min, D.; Ham, B.; Jeon, S.; Lin, S. & Sohn, K. (2017). Fcss: Fully convolutional self-similarity for dense semantic correspondence. *arXiv preprint arXiv:1702.00926*.

Kim, S.; Min, D.; Ham, B.; Ryu, S.; Do, M. N. & Sohn, K. (2015). Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2103--2112.

Kim, S.-W. & Oommen, B. J. (2004). On using prototype reduction schemes to optimize kernel-based nonlinear subspace methods. *Pattern Recognition*, 37(2):227--239.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755--1758.

Kohonen, T. (1988). Self-organization and associative memory. *Springer-Verlag Berlin Heidelberg New York. Also Springer Series in Information Sciences*, 8.

Krig, S. (2014). *Computer Vision Metrics: Survey, taxonomy, and analysis*. Apress, New York.

Krizhevsky, A. & Hinton, G. (2009). Learning multiple layers of features from tiny images. *Neurocomputing*.

Krizhevsky, A. & Hinton, G. (2010). Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40.

Krizhevsky, A.; Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097--1105.

Kuncheva, L. I. & Bezdek, J. C. (1998). An integrated framework for generalized nearest prototype classifier design. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 6(05):437--457.

Lake, B. M.; Zaremba, W.; Fergus, R. & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *CogSci*.

Lakoff, G. & Kövecses, Z. (1987). The cognitive model of anger inherent in american english. *Cultural models in language and thought*, pp. 195--221.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

LeCun, Y.; Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436.

LeCun, Y.; Boser, B. E.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. E. & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pp. 396--404.

Lecun, Y.; Bottou, L.; Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. ISSN 0018-9219.

LeCun, Y.; Bottou, L.; Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278--2324.

Lee, D. & Wang, X.-J. (2009). Mechanisms for stochastic decision making in the primate frontal cortex: Single-neuron recording and circuit modeling. In *Neuroeconomics*, pp. 481--501. Elsevier.

Lee, D. D. & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 556--562.

Leutenegger, S.; Chli, M. & Siegwart, R. Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2548--2555. IEEE.

Li, M. (2007). Texture moment for content-based image retrieval. In *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 508--511. IEEE.

Li, O.; Liu, H.; Chen, C. & Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lin, G.; Shen, C.; van den Hengel, A. & Reid, I. (2016a). Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3194--3203.

Lin, K.; Lu, J.; Chen, C.-S. & Zhou, J. (2016b). Learning compact binary descriptors with unsupervised deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1183--1192.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P. & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the of the European Conference on Computer Vision (ECCV)*, pp. 740--755. Springer.

Liu, C.; Yuen, J. & Torralba, A. (2011). Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(5):978--994.

Liu, S. & Deng, W. (2015). Very deep convolutional neural network based image classification using small training sample size. In *ACPR*, pp. 730--734. IEEE.

Long, J.; Shelhamer, E. & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431--3440.

Long, J. L.; Zhang, N. & Darrell, T. (2014). Do convnets learn correspondence? In *Advances in Neural Information Processing Systems*, pp. 1601--1609.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91--110.

Lu, X.; Yuan, Y. & Fang, J. (2017). Jm-net and cluster-svm for aerial scene classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2386--2392. AAAI Press.

Maaten, L. v. d. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579--2605.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pp. 281--297. Oakland, CA, USA.

Mainali, P.; Lafruit, G.; Tack, K.; Van Gool, L. & Lauwereins, R. (2014). Derivative-based scale invariant image feature detector with error resilience. *IEEE Transactions on Image Processing*, 23(5):2380--2391.

Malisiewicz, T. (2011). *Exemplar-based representations for object detection, association and beyond*. Carnegie Mellon University.

Malisiewicz, T. & Efros, A. A. (2008). Recognition by association via learning per-exemplar distances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1--8. IEEE.

Manen, S.; Guillaumin, M. & Van Gool, L. (2013). Prime object proposals with randomized prim's algorithm. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2536--2543.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, 58:25--45.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115--133.

McRae, K. & Jones, M. (2013). Semantic memory. *The Oxford handbook of Cognitive Psychology*, 206.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207.

Medin, D. L. & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35(1):113--138.

Mervis, C. B. & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32(1):89--115.

Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(10):1615--1630.

Minda, J. P. & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(3):775.

Minda, J. P. & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(2):275.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pp. 221--242. Springer.

Murphy, G. L. (2004). The big book of concepts. *Journal of child language*, 31(1):247--253.

Murthy, V. N.; Maji, S. & Manmatha, R. (2015). Automatic image annotation using deep learning representations. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 603--606. ACM.

Ng, A. Y. & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems*, pp. 841--848.

Nguyen, A.; Yosinski, J. & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427--436.

Nilsback, M.-E. & Zisserman, A. (2006). A visual vocabulary for flower classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 1447--1454. IEEE.

Nogueira, K.; Penatti, O. A. & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539--556.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39.

Ojala, T.; Pietikäinen, M. & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51--59.

Ojala, T.; Pietikainen, M. & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(7):971--987.

Okutomi, M. & Kanade, T. (1993). A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 15(4):353--363.

Oliva, A. (2016). 6.819/6.869. advances in computer vision: High-level vision.

Oliva, A. & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42(3):145--175.

O'Neill, S. P. (2015). Sapir–whorf hypothesis. *The International Encyclopedia of Language and Social Interaction*, pp. 1--10.

Osendorfer, C.; Bayer, J.; Urban, S. & Van Der Smagt, P. (2013). Convolutional neural networks learn compact local image descriptors. In *International Conference on Neural Information Processing*, pp. 624--630. Springer.

Perez, C. B. & Olague, G. (2013). Genetic programming as strategy for learning image descriptor operators. *Intelligent Data Analysis*, 17(4):561--583.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77(3,):353--363.

Qiu, W.; Wang, X.; Bai, X.; Yuille, A. & Tu, Z. (2014). Scale-space sift flow. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pp. 1112--1119. IEEE.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3(3):382--407.

Ren, S.; He, K.; Girshick, R. & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pp. 91--99.

Revaud, J.; Weinzaepfel, P.; Harchaoui, Z. & Schmid, C. (2015). Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1164--1172.

Rocco, I.; Arandjelovic, R. & Sivic, J. (2017). Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2.

Rocco, I.; Arandjelović, R. & Sivic, J. (2018). End-to-end weakly-supervised semantic alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rosch, E. (1973a). On the internal structure of perceptual and semantic categories. In *Cognitive Development and the Acquisition of Language*, pp. 111–144. Academic Pres, New York.

Rosch, E. (1975a). Cognitive reference points. *Cognitive Psychology*, 7(4):532--547.

Rosch, E. (1975b). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192.

Rosch, E. (1977). Human categorization. *Studies in cross-cultural psychology*, 1:1--49.

Rosch, E. (1978). Principles of categorization. In Rosch, E. & Lloyd, B. B., editores, *Cognition and Categorization*, volume 1, pp. 27–48. Lawrence Erlbaum Associates Hillsdale, NJ, Hillsdale, Michigan.

Rosch, E. (1988). Coherences and categorization: A historical view. *The development of language and language researchers: Essays in honor of Roger Brown*, pp. 373--392.

Rosch, E. & Lloyd, B. B. (1978). *Cognition and Categorization*, volume 1. Lawrence Erlbaum Associates Hillsdale, NJ.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive sychology*, 7(4):573--605.

Rosch, E.; Mervis, C. B.; Gray, W. D.; Johnson, D. M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3):382--439.

Rosch, E. H. (1973b). Natural categories. *Cognitive Psychology*, 4(3):328--350.

Rosenberg, A. & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386.

Rublee, E.; Rabaud, V.; Konolige, K. & Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2564--2571. IEEE.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C. & Fei-Fei, L. (2015a). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. et al. (2015b). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211--252.

Salakhutdinov, R. & Hinton, G. (2009). Deep boltzmann machines. In *Artificial Intelligence and Statistics*, pp. 448--455.

Saleh, B.; Elgammal, A. M. & Feldman, J. (2016). Incorporating prototype theory in convolutional neural networks. In *Proceedings of the IEEE International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 3446--3453.

Saw, J. G.; Yang, M. C. & Mo, T. C. (1984). Chebyshev inequality with estimated mean and variance. *The American Statistician*, 38(2):130--132.

Scharstein, D. & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision (IJCV)*, 47(1-3):7--42.

Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673--2681.

Seo, S. & Obermayer, K. (2003). Soft learning vector quantization. *Neural computation*, 15(7):1589--1604.

Sermanet, P.; Kavukcuoglu, K.; Chintala, S. & LeCun, Y. (2013). Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3626--3633.

Shechtman, E. & Irani, M. (2007). Matching local self-similarities across images and videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1--8. IEEE.

Shelhamer, E.; Long, J. & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(4):640--651.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317--1323.

Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P. & Moreno-Noguer, F. (2015). Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 118--126.

Simonyan, K.; Vedaldi, A. & Zisserman, A. (2014). Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8):1573--1585.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, E. E. & Medin, D. L. (1981). *Categories and concepts*, volume 9. Harvard University Press Cambridge, MA.

Snell, J.; Swersky, K. & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4080--4090.

Song, Y.-j.; Park, W.-b.; Kim, D.-w. & Ahn, J.-h. (2004). Content-based image retrieval using new color histogram. In *Intelligent Signal Processing and Communication Systems, 2004. ISPACS 2004. Proceedings of 2004 International Symposium on*, pp. 609--611. IEEE.

Stellato, B.; Van Parys, B. P. & Goulart, P. J. (2017). Multivariate chebyshev inequality with estimated mean and variance. *The American Statistician*, 71(2):123--127.

Sternberg, R. J. & Sternberg, K. (2016). *Cognitive Psychology*. Nelson Education.

Strecha, C.; Bronstein, A.; Bronstein, M. & Fua, P. (2012). Ldahash: Improved matching with smaller descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(1):66--78.

Strecha, C.; Von Hansen, W.; Van Gool, L.; Fua, P. & Thoennessen, U. (2008). On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1--8. Ieee.

Sun, C.; Shrivastava, A.; Singh, S. & Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *arXiv preprint arXiv:1707.02968*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V. & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pp. 4278--4284.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1--9.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J. & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818--2826.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Szeliski, R. (2006). Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1--104.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Taniai, T.; Sinha, S. N. & Sato, Y. (2016). Joint recovery of dense correspondence and cosegmentation in two images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4246--4255.

Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: inferring how from where. *Neuropsychologia*, 41(3):280--292.

Tola, E.; Lepetit, V. & Fua, P. (2008). A fast local descriptor for dense matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1--8. IEEE.

Trulls, E.; Kokkinos, I.; Sanfeliu, A. & Moreno-Noguer, F. (2013). Dense segmentation-aware descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2890--2897.

Trzcinski, T.; Christoudias, M.; Fua, P. & Lepetit, V. (2013). Boosting binary keypoint descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2874--2881.

Tulving, E. (1992). Memory systems and the brain. *Clinical neuropharmacology*, 15(Part A):327A--328A.

Tulving, E. (2007). Coding and representation: searching for a home in the brain. *Science of memory: Concepts*, pp. 65--68.

Uijlings, J. R.; Van De Sande, K. E.; Gevers, T. & Smeulders, A. W. (2013). Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154--171.

Ungerer, F. & Schmid, H.-J. (2013). *An introduction to cognitive linguistics*. Routledge.

Verdie, Y.; Yi, K.; Fua, P. & Lepetit, V. (2015). Tilde: a temporally invariant learned detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5279--5288.

Vinh, N. X.; Epps, J. & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *JMLR*, 11(Oct):2837--2854.

Waltz, D. L. (1975). Generating semantic descriptions from drawings of scenes with shadows. *The Phychology of Computer Vision*, pp. 19--92.

Wang, H.; Raj, B. & Xing, E. P. (2017). On the origin of deep learning. *arXiv preprint arXiv:1702.07800*.

Wang, K.; Wang, B. & Peng, L. (2009). Cvap: validation for cluster analyses. *Data Science Journal*, 8:88--93.

Werbos, P. (1975). *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Harvard University.

Winston, P. H. (1970). *Learning structural descriptions from examples.* Tese de doutorado, Massachusetts Institute of Technology (MIT).

Wittgenstein, L. (1953). Philosophical investigations, anscombe. *GEM (Trans.). Blackwell Publishing, Malden, MA*.

Wohlhart, P.; Köstinger, M.; Donoser, M.; Roth, P. M. & Bischof, H. (2013). Optimizing 1-nearest prototype classifiers. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 460--467. IEEE.

Xu, B.; Fu, Y.; Jiang, Y.-G.; Li, B. & Sigal, L. (2016). Video emotion recognition with transferred deep feature encodings. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pp. 15--22. ACM.

Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D. & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619--8624.

Yan, R. & Naphade, M. (2005). Semi-supervised cross feature learning for semantic concept detection in videos. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 657--663. IEEE.

Yang, H.; Lin, W.-Y. & Lu, J. (2014). Daisy filter flow: A generalized discrete approach to dense correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3406--3413.

Yang, J.; Parikh, D. & Batra, D. (2016). Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5147--5156.

Yi, K. M.; Trulls, E.; Lepetit, V. & Fua, P. (2016). Lift: Learned invariant feature transform. In *Proceedings of the of the European Conference on Computer Vision (ECCV)*, pp. 467--483. Springer.

Zabih, R. & Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. In *Proceedings of the of the European Conference on Computer Vision (ECCV)*, pp. 151--158. Springer.

Zagoruyko, S. & Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353--4361.

Zaki, S. R.; Nosofsky, R. M.; Stanton, R. D. & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment.

Journal of Experimental Psychology: Learning, Memory and Cognition, 29(6):1160--1173.

Zbontar, J. & LeCun, Y. (2015). Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1592--1599.

Zbontar, J. & LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the of the European Conference on Computer Vision (ECCV)*, pp. 818--833. Springer.

Zhang, Z.; Sturgess, P.; Sengupta, S.; Crook, N. & Torr, P. H. (2012). Efficient discriminative learning of parametric nearest neighbor classifiers. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2232--2239. IEEE.

Zhao, H. & Qin, Z. (2015). Clustering data and vague concepts using prototype theory interpreted label semantics. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pp. 236--246. Springer.

Zhou, T.; Krahenbuhl, P.; Aubry, M.; Huang, Q. & Efros, A. A. (2016). Learning dense correspondence via 3d-guided cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 117--126.

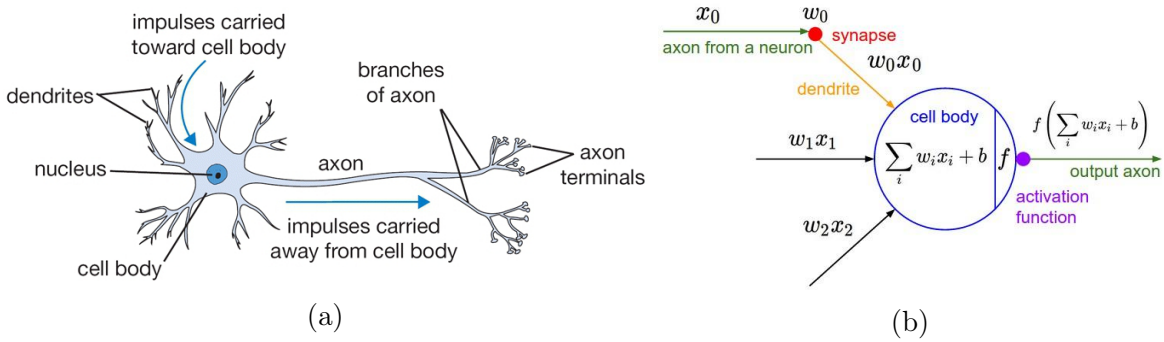
Zhu, L.; Shen, J.; Xie, L. & Cheng, Z. (2017). Unsupervised visual hashing with semantic assistant for content-based image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 29(2):472--486.

Zitnick, C. L. & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *Proceedings of the of the European Conference on Computer Vision (ECCV)*, pp. 391--405. Springer.

Apêndice A

Redes Neuronais Convolucionais

As Redes Neuronais (RNs)(*Neural Networks (NNs)*) originalmente foram inspiradas nas conexões biológicas do cérebro humano. Desde seus primórdios, as RNs tiveram como propósito fundamental criar um método de aprendizagem baseado na modelagem dos sistemas neuronais biológicos (Figura A.1a). Os primeiros trabalhos nessa área remetem-se ao século XX, cujos pressupostos teóricos definiram e estabeleceram os conceitos fundamentais desse paradigma (que sempre esteve restrito pelas limitações de hardware). De maneira geral, no funcionamento de uma RN cada neurônio recebe algumas entradas, realiza um produto escalar e em seguida aplica uma função de ativação (Figura A.1b).



(a) (b)

Figura A.1: As Redes Neuronais originalmente foram inspiradas nas conexões biológicas do cérebro humano. a) neurônio biológico, b) modelo matemático comum de uma RN. Fonte: Karpathy, 2017.

Os principais avanços no percurso da evolução das RNs foram manifestados em três momentos diferentes. Entre os anos 1940-1960 surgiram os primeiros modelos de RNs, e os trabalhos de McCulloch & Pitts (1943); Hebb (1949); Rosenblatt (1958) foram os mais destacados nesse período. Somente 20 anos depois, os estudos de Werbos

(1975); Fukushima & Miyake (1982); Ackley et al. (1985); Jordan (1986) revolucionaram a área novamente introduzindo conceitos como *backpropagation*, *neocognitron* (inspiração das CNNs atuais), *boltzmann machines* e *Recurrent Neural Network (RNN)* respectivamente. Na década de 1990, LeCun et al. (1990) apresentaram com a rede LeNet a possibilidade do uso das redes profundas na prática. Nesse mesmo período, os estudos de Schuster & Paliwal (1997) e Hochreiter & Schmidhuber (1997) aperfeiçoaram o trabalho de Jordan (1986) com a introdução das RNNs bidirecionais (*Bidirectional Recurrent Neural Networks (BRNNs)*) e com a arquitetura *Long Short-Term Memory (LSTM)*. Esses trabalhos mostraram que somente quando existiram avanços tecnológicos significativos no hardware, a área de Aprendizagem Profundo (CNNs) retornaria a ser foco de atenção pela comunidade acadêmica.

As RNs constituem um paradigma que — diferente do enfoque convencional— não define explicitamente como resolver um determinado problema ou subproblema. Ao contrário, as RNs aprendem dos dados observacionais e calculam uma solução para o problema analisado.

Praticamente até 2006, era desconhecido como treinar as RNs para superarem os enfoques tradicionais (exceto para alguns problemas especializados), fato que mudou com o surgimento das técnicas de aprendizagem das RNs profundas. Depois de 2006 trabalhos como os de Hinton et al. (2006), Salakhutdinov & Hinton (2009) e Hinton et al. (2012) apresentaram novas técnicas que abriram a atual era da aprendizagem profunda, cuja aplicação permitiu um rendimento destacado na resolução de tarefas de Visão Computacional, de Reconhecimento de Voz e de Processamento da Linguagem Natural.

A.1 Primórdios das CNNs

As origens das Redes Neurais Convolucionais (*CNNs*) remontam-se à década de 1980 com o estudo de Fukushima. LeCun et al. (1998) apresentaram o documento seminal (com a rede *LeNet*) que estabelece o tópico moderno das CNNs. A rede LeNet é conhecida pela capacidade de classificar dígitos e lidar com uma ampla variedade de problemas diferentes na imagem, incluindo variações de escala, de rotação, etc. Juntamente com a introdução da rede LeNet, LeCun (1998) também apresentou o banco de dados MNIST, atualmente o banco de dados utilizado como (*benchmark*) na área de reconhecimento de dígitos.

As CNNs constituem uma família dentro do conjunto de modelos da aprendizagem profunda. Esse tipo de rede, apesar de possuir similitudes com as RNs comuns, possui

uma inspiração biológica diferente. As CNNs evoluíram a partir do conhecimento do córtex visual humano (Hubel & Wiesel, 1959). O desenho biônico (*bionic design*)[1] das CNNs na réplica do sistema visual humano representa a causa do sucesso atual em tarefas de Visão Computacional (Wang et al., 2017).

Apesar das CNNs possuirem diferenças com respeito às RNs, elas também usam técnicas comuns como *backpropagation*, *gradient descent*, *regularization* e *non-linear activation functions*. O que é diferente? A arquitetura da CNN supõe explicitamente que as entradas são imagens, característica que permite codificar certas propriedades na arquitetura. Essa característica torna a função direta (*forward function*) mais eficiente na programação, reduzindo a quantidade de parâmetros da rede (Karpathy, 2017).

Com o sucesso da LeNet foi constatada a capacidade das CNNs na realização bem sucedida de tarefas de Visão Computacional. As potencialidades das CNNs atraíram o interesse da comunidade acadêmica para resolver o problema de reconhecimento de objetos na tarefa de classificação de CIFAR (Krizhevsky & Hinton, 2010) e no desafio da ImageNet (*ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*) (Russakovsky et al., 2015b). Foi com o desafio de ImageNet que apareceram os modelos mais representativos da família CNN. O modelo *AlexNet* (Krizhevsky et al., 2012) marcou o ponto de inflexão nesse desafio, pois pela primeira vez uma rede CNN profunda ganhou aquela competição. O recorde estabelecido pelo modelo de Krizhevsky et al. (2012) (ao reduzir drasticamente o erro de classificação de 26% para 15%) motivou que muitas empresas começassem usar intensamente a aprendizagem profunda nos seus serviços (Ex: Google, Facebook, Amazon, Pinterest, Instagram, etc).

A.2 Estrutura e conceitos

As CNNs podem ser definidas a *grosso modo*, como uma rede neuronal hierárquica multicamada. As CNNs constituem redes especializadas no processamento de imagens que podem aprender as relações entre entradas e saídas (Wang et al., 2017). A suposição explícita de que as entradas são imagens permite codificar certas propriedades na arquitetura, visando ganhar em eficiência e na redução da quantidade de parâmetros na rede. Na obtenção desses resultados, as CNN exploram a operação convolução, propriedade que define a denominação da arquitetura desse tipo de rede.

A operação de convolução é utilizada frequentemente em diversas tarefas de Visão Computacional. Essa operação consiste em "filtrar" uma imagem usando uma más-

[1] desenho biônico refere-se ao tipo de construção onde o modelo (ou produto) é composto por características e estruturas que representam "substituições" de estruturas e processos anatômicos de inspiração biológica.

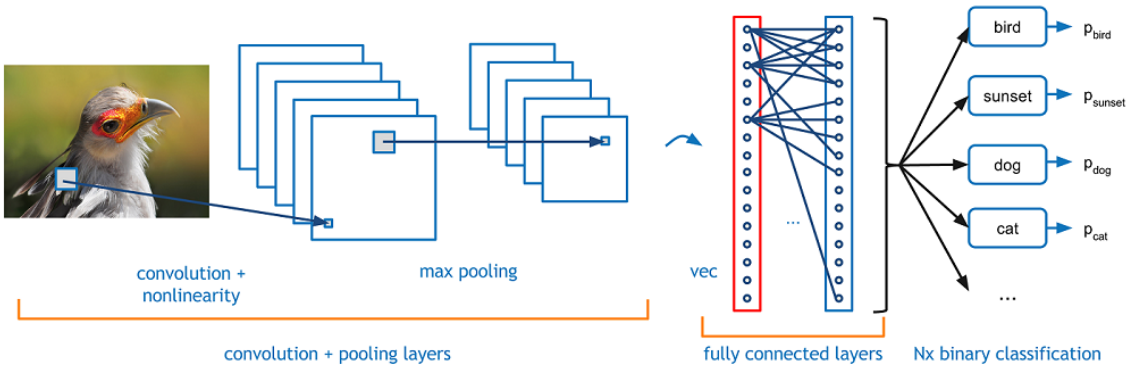


Figura A.2: Arquitetura geral das Redes Neuronais Convolucionais. Fonte: Deshpande, 2017.

cara (conjuntos de valores chamados pesos) produzindo um pixel de saída que resulta em uma combinação linear dos píxeis de entrada com os pesos da máscara. Diferentes máscaras produzem distintos resultados na operação convolução, e representam —nas CNNs— a conectividade entre as camadas sucessivas. Essa organização explorada pelas CNNs justifica —junto com outras características— o sucesso alcançado por esse tipo de arquitetura (Wang et al., 2017).

As CNNs usam de forma consecutiva pequenos fragmentos de informação com o propósito de combinar informação nas camadas (*layers*) mais profundas. A ideia principal consiste em realizar operações sucessivas mudando a complexidade da máscara (máscaras simples detectarão padrões simples e outras de maior complexidade encontrarão padrões mais complexos). Com esse procedimento podem ser encontradas bordas, quinas, etc. na primeira camada. Em camadas posteriores são combinadas essas características simples para transformá-las em padrões que descrevem a posição do objeto, iluminação, escalas, etc. As camadas finais tentam corresponder a imagem de entrada com todos os padrões encontrados, produzindo uma predição final que é resultado da soma ponderada de todos os padrões. Essa estrutura fornece às CNNs a capacidade de modelar complexas variações e comportamentos, provendo predições muito precisas.

Uma camada de uma CNNs constitui-se em um volume tridimensional de neurônios: altura, largura e profundidade. Geralmente, essas redes são construídas com uma estrutura que contêm três tipos distintos de camadas (Ver Figura A.2):

- Camada convolucional (*convolutional layer*): requer o uso de máscaras (*kernels*). Realiza filtragem usando diferentes máscaras e, em conjunto com uma função de não linearidade, produz os mapas de características (*feature maps*). A convolução aproveita as ideias de: interações dispersas, parâmetros comparti-

lhados e representações equivariantes, visando melhorar a eficiência do sistema e reduzir drasticamente a quantidade de parâmetros da rede.

- Camada de redução (*pooling layer*): geralmente são usadas imediatamente após as camadas convolucionais. A utilidade principal dessas camadas reside na redução das dimensões espaciais (largura × altura) do volume de entrada para a seguinte camada convolucional. A operação realizada nessa camada é denominada redução de amostragem, pois a redução do tamanho conduz à perda proveitosa de informação (Wang et al., 2017).

- Camada totalmente conectada (*fully-connected layer*): de forma geral essa camada é usada como a última camada da estrutura da rede. Nessa camada perde-se a informação espacial. Normalmente é usada como camada classificadora, cujo número de neurônios é o mesmo que a quantidade de classes que são preditas.

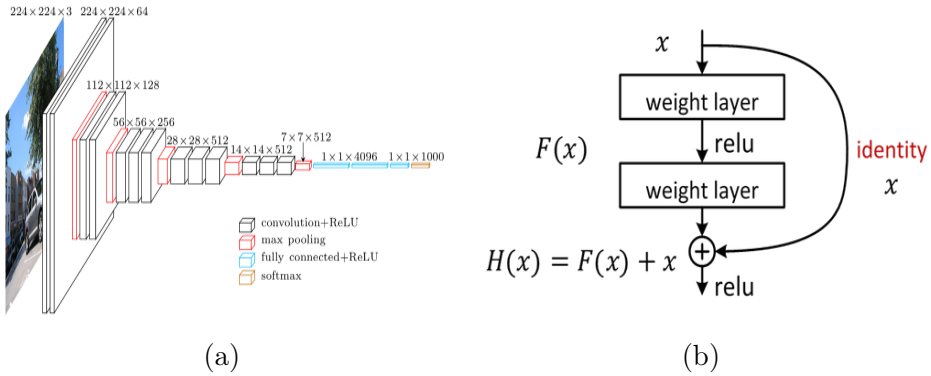
Existem duas arquiteturas básicas de CNN: a CNN e as Redes Completamente Convolucionais (*Fully Convolutional Network (FCN)*). A arquitetura das CNNs produz, para toda a imagem, uma saída do tipo "todas conectadas com todas"(*fully connected*). Aliás, a arquitetura FCN possui um codificador e um decodificador, comprime a informação e entrega normalmente um pixel de saída por cada pixel de entrada.

A.3 Modelos relevantes

A generalização da representação de caraterísticas aprendidas pelos modelos CNNs, em grandes volumes de dados, permitiu desenvolver outras tarefas como detecção de objetos (Girshick et al., 2014; Sermanet et al., 2013), classificação de cenas (Donahue et al., 2014) e segmentação (Lin et al., 2016a).

A solução proposta por Krizhevsky et al. (2012) no desafio de ImageNet, motivou uma evolução acelerada das CNNs. Desenvolveram-se diversas propostas nos anos seguintes que permitiram entender e visualizar o funcionamento interno desses sistemas. Zeiler & Fergus (2014) apresentaram um dos maiores aporte teóricos com a arquitetura da rede ZFNet. Essa arquitetura constitui uma otimização da estrutura da rede AlexNet e suas principais melhorias residem no uso da função de ativação não linear (*Rectified Linear Units (ReLU)*) na perda de entropia cruzada (*cross-entropy loss*) como função de erro e no treinamento em lote do gradiente estocástico descendente (*batch stochastic gradient descente*).

Os mesmos autores, desenvolveram uma técnica de visualização denominada Rede Deconvolucional (*Deconvolutional Network - deconvnet*) que, diferente do que faz uma



(a) (b)

Figura A.3: Exemplos de modelos CNN para tarefas de classificação. a) Arquitetura do modelo VGG b) Arquitetura do modelo ResNet. Fonte: Simonyan & Zisserman, 2014; He et al., 2016 respectivamente.

camada convolucional, permite examinar diferentes ativações de características e mapear a relação com o espaço de entrada (*píxeis*). A ZFNet proveu uma importante intuição quanto ao funcionamento das CNNs, e ilustrou novas formas de melhorar o desempenho das CNN. A abordagem de visualização pelo uso da *deconvnet* e as demais melhorias propostas, forneceram à comunidade científica ferramentas potentes para a evolução das CNNs.

A simplicidade e a profundidade foram as principais características do modelo apresentado por Simonyan & Zisserman (2014), denominado VGG (Ver Figura A.3a). O modelo proposto constituiu o mais profundo da época (19 camadas), mas a sua arquitetura era muito simples. Na arquitetura da VGG todas as camadas possuem uma camada convolucional 3x3 e uma camada de agrupamento 2x2 (*pooling*). Esse simples uso da camada convolucional simula um filtro maior, mantendo os benefícios de filtros de tamanhos menores. Por exemplo, a combinação de duas camadas convolucionais de 3x3 tem um campo receptivo efetivo de uma camada de 5x5, mas com menos parâmetros.

Na arquitetura do modelo VGG o tamanho espacial dos volumes de entrada em cada camada diminui como resultado das camadas de convolução e de agrupamento. Não obstante, a profundidade dos volumes aumenta devido ao aumento do número de filtros usados (especificamente o número de filtros é duplicado após cada camada de agrupamento). Esse comportamento reforça a ideia da arquitetura de diminuir as dimensões espaciais, mas aumentando em profundidade (Ver Figura A.3a.)

A rede VGG foi projetada para o desafio de ImageNet 2015, mas não foi a vencedora da competição naquele ano (o vencedor foi GoogLeNet (Szegedy et al., 2015)). A rede GoogLeNet introduziu vários conceitos importantes como o módulo de *Inception*,

e esse conceito principal foi utilizado posteriormente pela rede R-CNN (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015). O design arbitrário/criativo da arquitetura da GoogLeNet apenas contribuiu à sociedade científica em comparação com os fundamentos teóricos instituídos pela rede VGG. A rede residual ResNet (He et al., 2016) corroborou os pressupostos anteriores, pois posteriormente, seguindo a abordagem da rede VGG, venceu o desafio ImageNet em um nível sem precedentes.

A *Microsoft Research Asia* apoiou-se na ideia de simplicidade e de profundidade de Simonyan & Zisserman (2014) para a construção da arquitetura da rede Residual Net (ResNet). He et al. (2016) projetaram com 152 camadas a rede ResNet, sendo dez vezes mais profunda que as redes da época. A arquitetura incrível de ResNet estabeleceu, além do recorde de profundidade, registros excelentes em tarefas de classificação, detecção e localização. A ResNet ganhou em 2015 o desafio ILSVRC com uma taxa de erro sem precedentes de 3.6 %. Tendo em conta que, dependendo da habilidade e da experiência, os seres humanos geralmente permanecem em torno a uma taxa de erro de 5-10 %, o resultado alcançado pela abordagem de He et al. (2016) marcou o início de uma nova era na área da aprendizagem profunda.

A ResNet demonstrou que o paradigma das redes convolucionais, além de ganhar em desempenho às abordagens tradicionais, pode ultrapassar em desempenho as habilidades humanas; uma afirmação que pode surpreender ao mais cético. He et al. (2016) usaram, além da abordagem "ultra-profunda", uma estrutura chamada bloque residual (*residual block*). A abordagem do bloque residual (Ver Figura A.3b) não usa a abordagem clássica de calcular a transformação direta da entrada (x) para a saída F(x). Os autores propuseram calcular uma leve alteração ou *delta* (a antiga saída F(x)) para obter uma nova representação ligeiramente alterada H(x). He et al. (2016) mostraram que era mais fácil otimizar o mapeamento residual do que otimizar o mapeamento original não referenciado.

Depois dos resultados alcançados por ResNet parecia praticamente impossível melhorar a taxa de erro em tarefas de classificação, detecção e localização. Porém, as versões de ResNet dos anos seguintes, associadas a outras abordagens, reduziram ainda mais essa taxa de erro.

A análise desenvolvida por Sun et al. (2017)(Figura A.4) mostra como desde 2012 houve avanços significativos nas capacidades de representação desses sistemas para tarefas de Visão Computacional. Esses progressos são consequência — e possuem uma correlação forte — do aumento do poder computacional (principalmente das GPUs), do aumento da complexidade e profundidade dos modelos projetados, e da disponibilidade dos dados rotulados em grande escala. Nesse sentido, os autores destacam como evoluíram as técnicas e o hardware enquanto, surpreendentemente, o tamanho

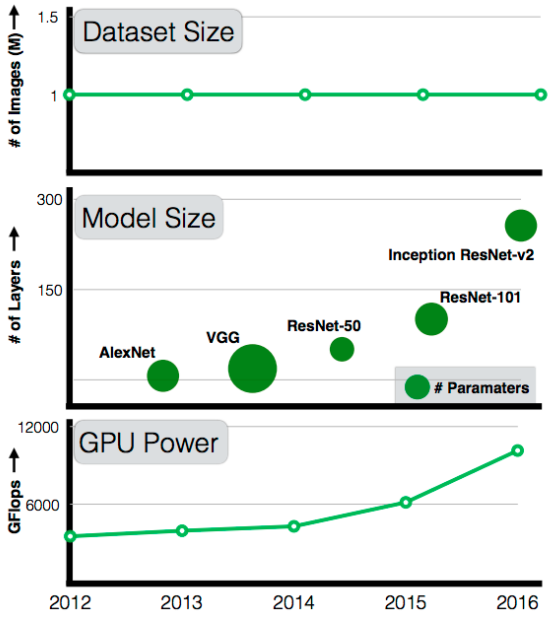


Figura A.4: Evolução das Redes Neuronais Convolucionais. Observa-se o aumentou do poder de processamento (GPU) e dos tamanhos dos modelos usados ao longo dos últimos anos, enquanto o tamanho do maior conjunto de dados de treinamento permaneceu constante. Fonte: Sun et al., 2017.

do maior conjunto de dados de treinamento (*ImageNet*) permaneceu constante.

Os trabalhos anteriores evidenciam o excelente desempenho das CNNs para resolver tarefas de Visão Computacional. O estado da arte atual, mostra que a maioria das tarefas de visão é resolvida por esses sistemas, corroborando o sucesso das CNNs como paradigma dominante.

O poder das CNNs para a extração de características abstratas e discriminativas, propiciou seu uso em tarefas que envolvem processamento da informação semântica. Vários trabalhos exploram as potencialidades das CNNs em tarefas de processamento semântico como a classificação semântica de imagens (Krizhevsky et al., 2012; Erhan et al., 2014) e a segmentação semântica (Chen et al., 2016; Shelhamer et al., 2017). Esses trabalhos justificam que explorar soluções de extração/representação de características com informação semântica usando CNNs, é uma opção admissível e bem sustentada.

Apêndice B

Arquitetura dos Modelos CNN de classificação usados

B.1 simples-MNIST

(Baseado na arquitetura da rede *LeNet* (Lecun et al., 1998))

```
_________________________________________________________________
Model Name: simples-MNIST
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_17 (Conv2D)           (None, 26, 26, 32)        320
_________________________________________________________________
conv2d_18 (Conv2D)           (None, 24, 24, 64)        18496
_________________________________________________________________
max_pooling2d_9 (MaxPooling2 (None, 12, 12, 64)        0
_________________________________________________________________
dropout_9 (Dropout)          (None, 12, 12, 64)        0
_________________________________________________________________
flatten_9 (Flatten)          (None, 9216)              0
_________________________________________________________________
features (Dense)             (None, 128)               1179776
_________________________________________________________________
featuresD (Dropout)          (None, 128)               0
_________________________________________________________________
last (Dense)                 (None, 10)                1280
=================================================================
Total params: 1,199,872
Trainable params: 1,199,872
Non-trainable params: 0
```

B.2 simples-CIFAR10

(Baseado na arquitetura da rede *Deep Belief Network* (Krizhevsky & Hinton, 2010))

```
_________________________________________________________________
Model Name: simples-CIFAR10
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_19 (Conv2D)           (None, 32, 32, 32)        896
_________________________________________________________________
activation_1 (Activation)    (None, 32, 32, 32)        0
_________________________________________________________________
conv2d_20 (Conv2D)           (None, 30, 30, 32)        9248
_________________________________________________________________
activation_2 (Activation)    (None, 30, 30, 32)        0
_________________________________________________________________
max_pooling2d_10 (MaxPooling (None, 15, 15, 32)        0
_________________________________________________________________
dropout_10 (Dropout)         (None, 15, 15, 32)        0
_________________________________________________________________
conv2d_21 (Conv2D)           (None, 15, 15, 64)        18496
_________________________________________________________________
activation_3 (Activation)    (None, 15, 15, 64)        0
_________________________________________________________________
conv2d_22 (Conv2D)           (None, 13, 13, 64)        36928
_________________________________________________________________
activation_4 (Activation)    (None, 13, 13, 64)        0
_________________________________________________________________
max_pooling2d_11 (MaxPooling (None, 6, 6, 64)          0
_________________________________________________________________
dropout_11 (Dropout)         (None, 6, 6, 64)          0
_________________________________________________________________
flatten_10 (Flatten)         (None, 2304)              0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               1180160
_________________________________________________________________
activation_5 (Activation)    (None, 512)               0
_________________________________________________________________
featuresD (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 10)                5120
_________________________________________________________________
activation_6 (Activation)    (None, 10)                0
=================================================================
Total params: 1,250,848
Trainable params: 1,250,848
Non-trainable params: 0
```

B.3 simples-CIFAR100

(Baseado na arquitetura da rede *Deep Belief Network* (Krizhevsky & Hinton, 2010))

```
_________________________________________________________________
Model Name: simples-CIFAR100
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 32, 32, 32)        896
_________________________________________________________________
activation_1 (Activation)    (None, 32, 32, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 30, 30, 32)        9248
_________________________________________________________________
activation_2 (Activation)    (None, 30, 30, 32)        0
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 15, 15, 32)        0
_________________________________________________________________
dropout_1 (Dropout)          (None, 15, 15, 32)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 15, 15, 64)        18496
_________________________________________________________________
activation_3 (Activation)    (None, 15, 15, 64)        0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 13, 13, 64)        36928
_________________________________________________________________
activation_4 (Activation)    (None, 13, 13, 64)        0
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 6, 6, 64)          0
_________________________________________________________________
dropout_2 (Dropout)          (None, 6, 6, 64)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 2304)              0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               1180160
_________________________________________________________________
activation_5 (Activation)    (None, 512)               0
_________________________________________________________________
featuresD (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 100)               51300
_________________________________________________________________
activation_6 (Activation)    (None, 100)               0
=================================================================
Total params: 1,297,028
Trainable params: 1,297,028
Non-trainable params: 0
```

B.4 VGG-CIFAR10

```
=================================================================
Model Name: VGG_CIFAR10
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 32, 32, 64)        1792
_________________________________________________________________
activation_1 (Activation)    (None, 32, 32, 64)        0
_________________________________________________________________
batch_normalization_1 (Batch (None, 32, 32, 64)        256
_________________________________________________________________
dropout_1 (Dropout)          (None, 32, 32, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 32, 32, 64)        36928
_________________________________________________________________
activation_2 (Activation)    (None, 32, 32, 64)        0
_________________________________________________________________
batch_normalization_2 (Batch (None, 32, 32, 64)        256
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 16, 16, 64)        0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 16, 16, 128)       73856
_________________________________________________________________
activation_3 (Activation)    (None, 16, 16, 128)       0
_________________________________________________________________
batch_normalization_3 (Batch (None, 16, 16, 128)       512
_________________________________________________________________
dropout_2 (Dropout)          (None, 16, 16, 128)       0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 16, 16, 128)       147584
_________________________________________________________________
activation_4 (Activation)    (None, 16, 16, 128)       0
_________________________________________________________________
batch_normalization_4 (Batch (None, 16, 16, 128)       512
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 8, 8, 128)         0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 8, 8, 256)         295168
_________________________________________________________________
activation_5 (Activation)    (None, 8, 8, 256)         0
_________________________________________________________________
batch_normalization_5 (Batch (None, 8, 8, 256)         1024
_________________________________________________________________
dropout_3 (Dropout)          (None, 8, 8, 256)         0
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 8, 8, 256)         590080
_________________________________________________________________
activation_6 (Activation)    (None, 8, 8, 256)         0
_________________________________________________________________
batch_normalization_6 (Batch (None, 8, 8, 256)         1024
_________________________________________________________________
dropout_4 (Dropout)          (None, 8, 8, 256)         0
_________________________________________________________________
```

```
conv2d_7 (Conv2D)            (None, 8, 8, 256)         590080
_________________________________________________________________
activation_7 (Activation)    (None, 8, 8, 256)         0
_________________________________________________________________
batch_normalization_7 (Batch (None, 8, 8, 256)         1024
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 4, 4, 256)         0
_________________________________________________________________
conv2d_8 (Conv2D)            (None, 4, 4, 512)         1180160
_________________________________________________________________
activation_8 (Activation)    (None, 4, 4, 512)         0
_________________________________________________________________
batch_normalization_8 (Batch (None, 4, 4, 512)         2048
_________________________________________________________________
dropout_5 (Dropout)          (None, 4, 4, 512)         0
_________________________________________________________________
conv2d_9 (Conv2D)            (None, 4, 4, 512)         2359808
_________________________________________________________________
activation_9 (Activation)    (None, 4, 4, 512)         0
_________________________________________________________________
batch_normalization_9 (Batch (None, 4, 4, 512)         2048
_________________________________________________________________
dropout_6 (Dropout)          (None, 4, 4, 512)         0
_________________________________________________________________
conv2d_10 (Conv2D)           (None, 4, 4, 512)         2359808
_________________________________________________________________
activation_10 (Activation)   (None, 4, 4, 512)         0
_________________________________________________________________
batch_normalization_10 (Batc (None, 4, 4, 512)         2048
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 2, 2, 512)         0
_________________________________________________________________
conv2d_11 (Conv2D)           (None, 2, 2, 512)         2359808
_________________________________________________________________
activation_11 (Activation)   (None, 2, 2, 512)         0
_________________________________________________________________
batch_normalization_11 (Batc (None, 2, 2, 512)         2048
_________________________________________________________________
dropout_7 (Dropout)          (None, 2, 2, 512)         0
_________________________________________________________________
conv2d_12 (Conv2D)           (None, 2, 2, 512)         2359808
_________________________________________________________________
activation_12 (Activation)   (None, 2, 2, 512)         0
_________________________________________________________________
batch_normalization_12 (Batc (None, 2, 2, 512)         2048
_________________________________________________________________
dropout_8 (Dropout)          (None, 2, 2, 512)         0
_________________________________________________________________
conv2d_13 (Conv2D)           (None, 2, 2, 512)         2359808
_________________________________________________________________
activation_13 (Activation)   (None, 2, 2, 512)         0
_________________________________________________________________
batch_normalization_13 (Batc (None, 2, 2, 512)         2048
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 1, 1, 512)         0
```

```
_________________________________________________________________
dropout_9 (Dropout)          (None, 1, 1, 512)         0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 512)               262656
_________________________________________________________________
activation_14 (Activation)   (None, 512)               0
_________________________________________________________________
batch_normalization_14 (Batc (None, 512)               2048
_________________________________________________________________
dropout_10 (Dropout)         (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 10)                5130
_________________________________________________________________
activation_15 (Activation)   (None, 10)                0
=================================================================
Total params: 15,001,418
Trainable params: 14,991,946
Non-trainable params: 9,472
```

Apêndice C

Exemplos de Protótipos Construídos

C.1 Protótipos no banco de dados MNIST

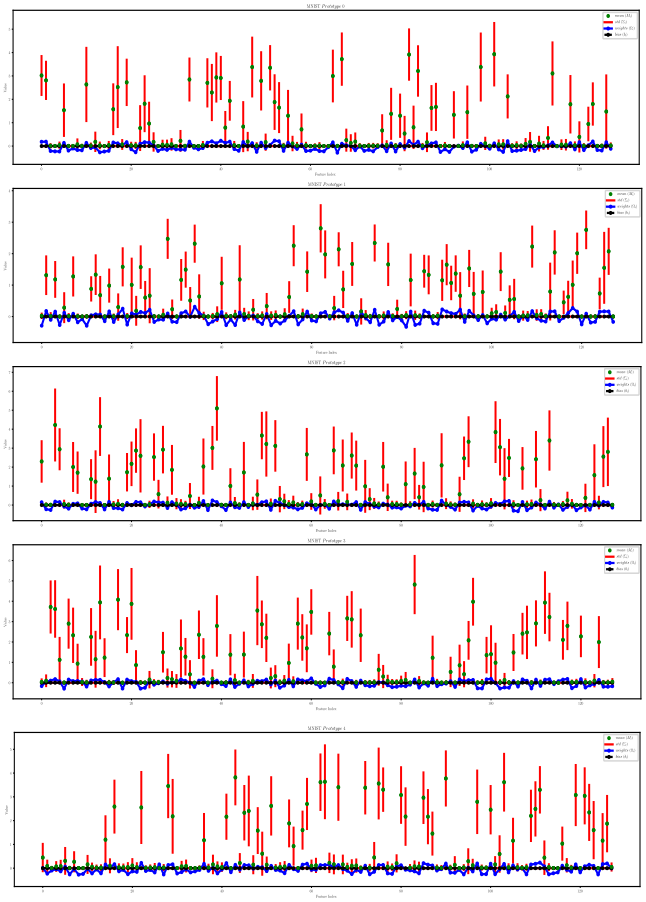


Figura C.1: Representação gráfica dos protótipos calculados para as 5 primeiras categorias do banco de dados MNIST usando o modelo *simples-MNIST*. Apresenta-se em cada representação (de 128 dimensões), o intervalo dos valores esperados das características unitárias da categoria, além dos valores de relevância correspondentes.

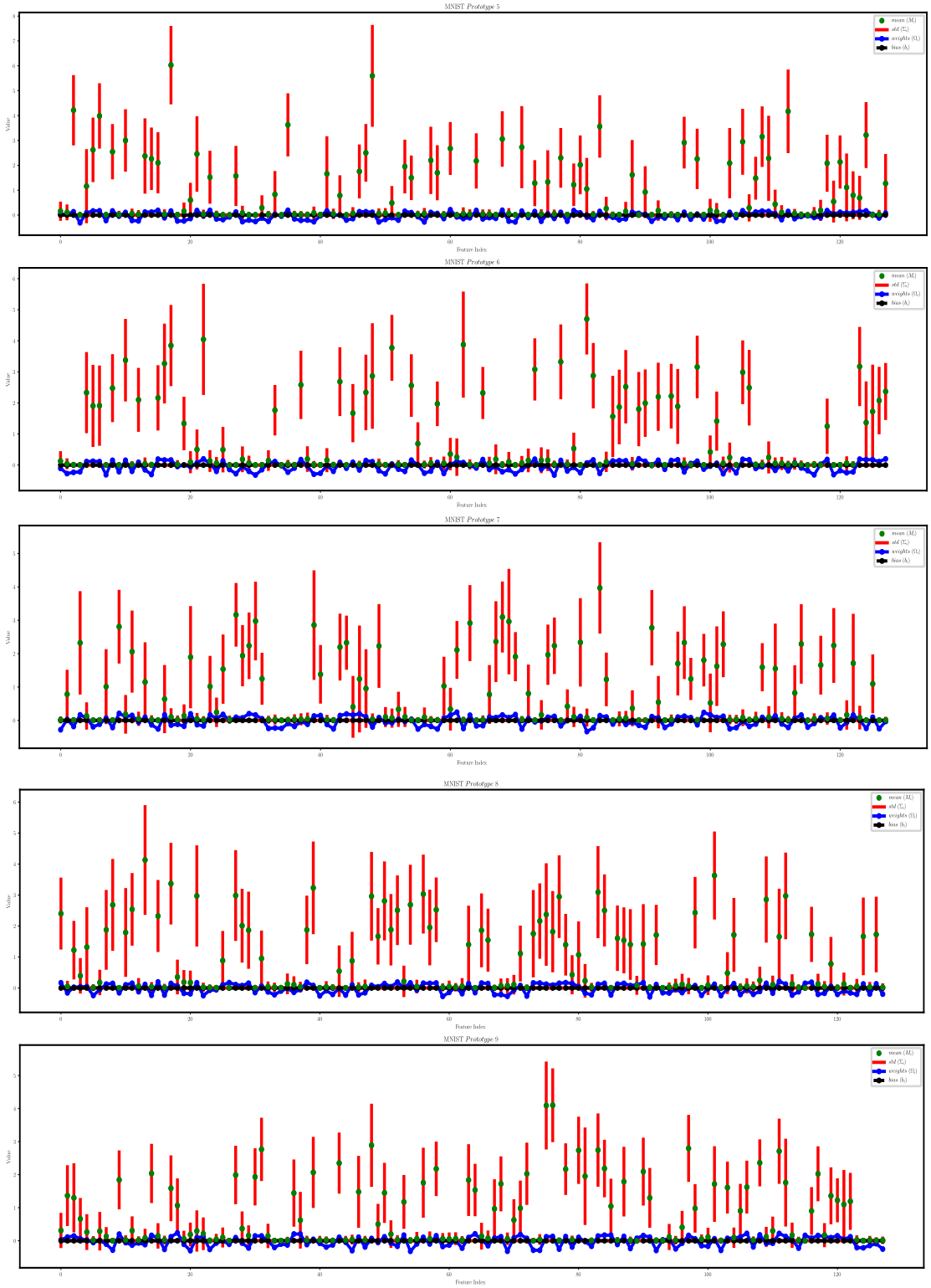


Figura C.2: Representação gráfica dos protótipos calculados para as 5 últimas categorias do banco de dados MNIST usando o modelo *simples-MNIST*.

C.2 Protótipos no banco de dados CIFAR10

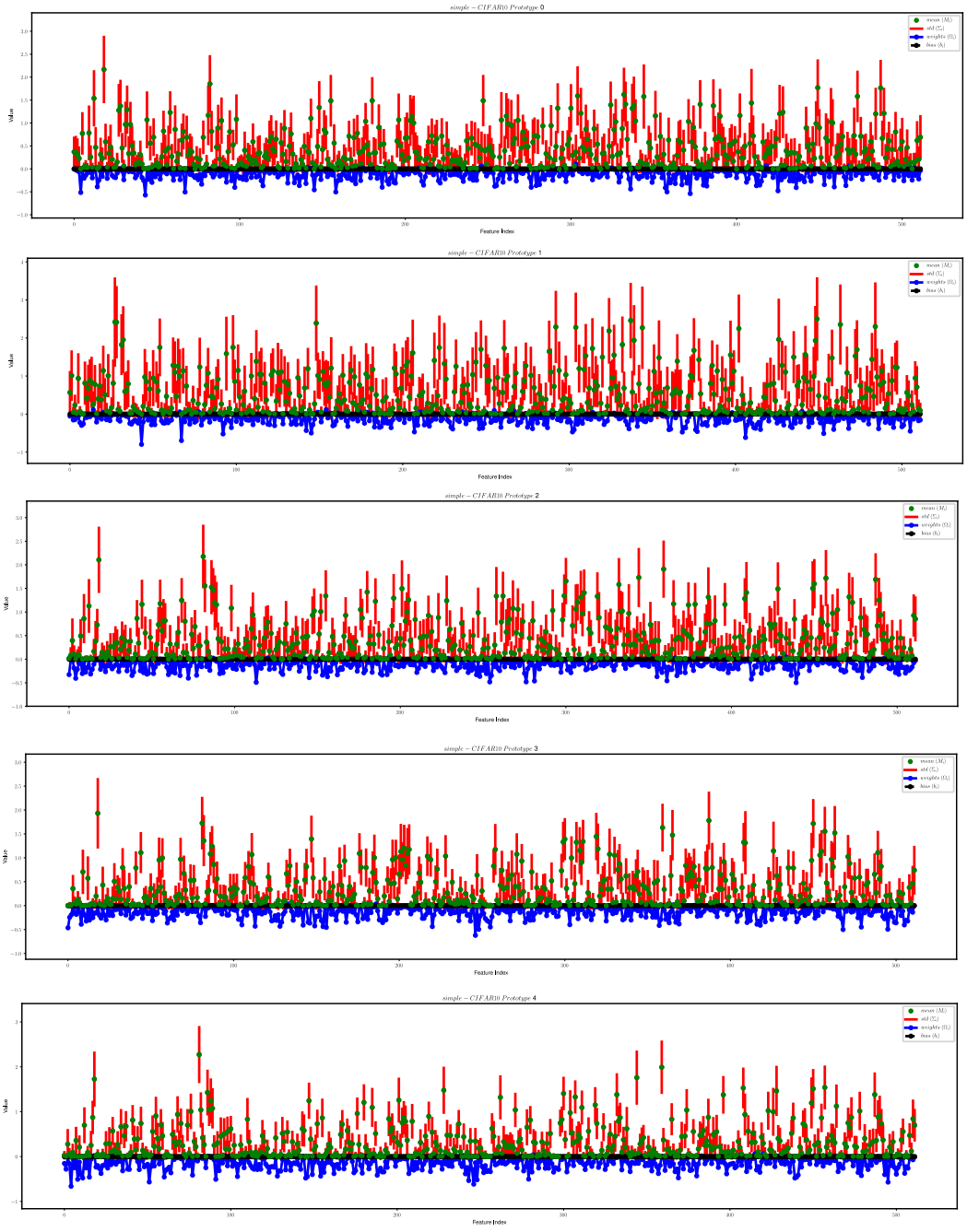


Figura C.3: Representação gráfica dos protótipos calculados para as 5 primeiras categorias do banco de dados CIFAR10 usando o modelo *simples-CIFAR10*. Apresenta-se em cada representação (de 512 dimensões), o intervalo dos valores esperados das características unitárias da categoria, além dos valores de relevância correspondentes.

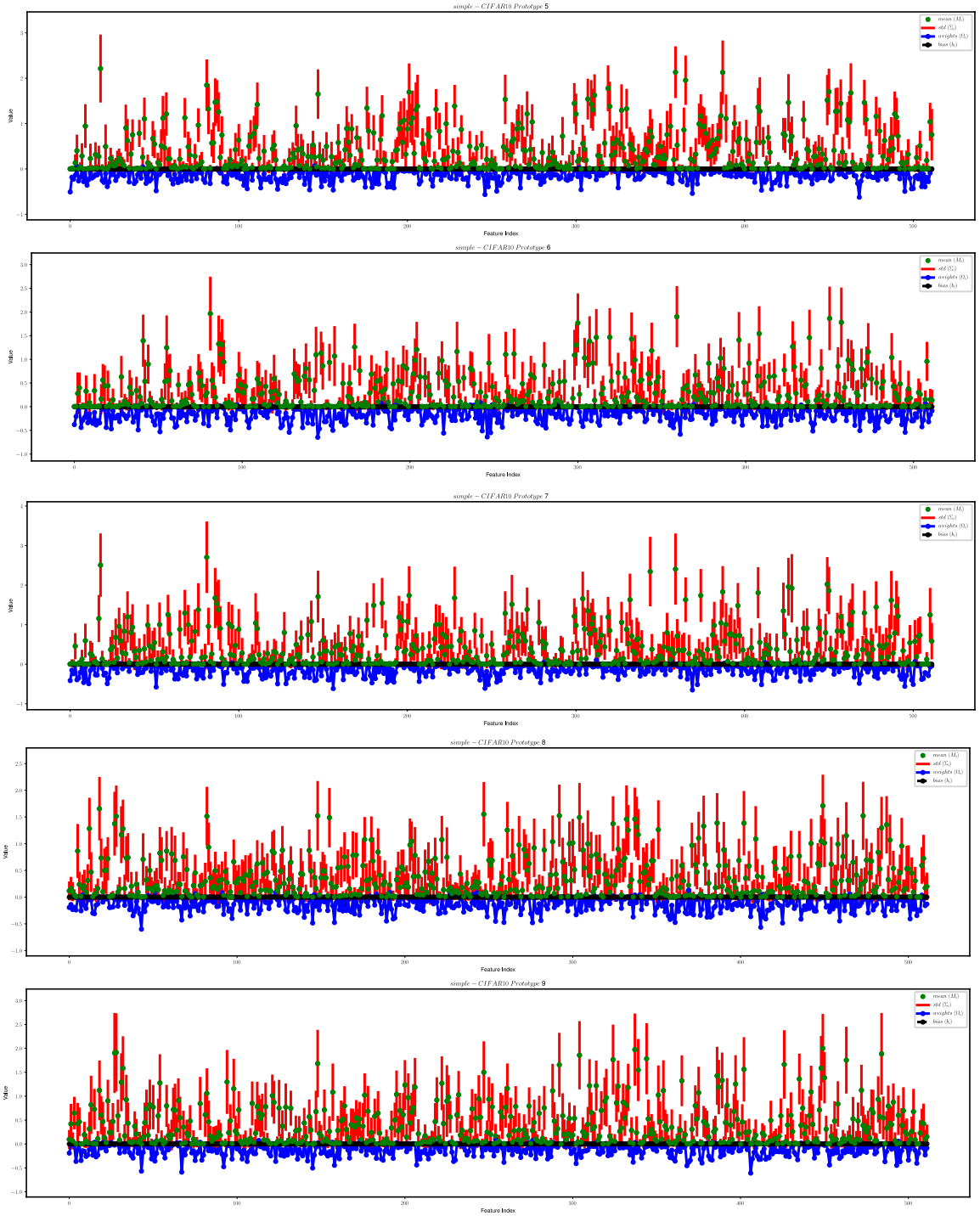
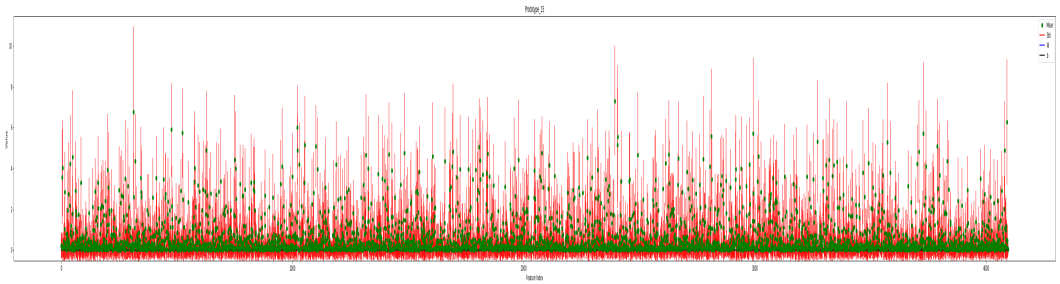
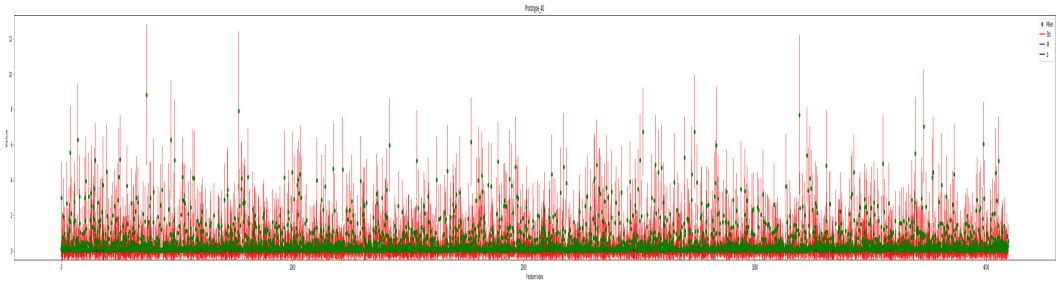


Figura C.4: Representação gráfica dos protótipos calculados para as 5 últimas categorias do banco de dados CIFAR10 usando o modelo *simples-CIFAR10*.

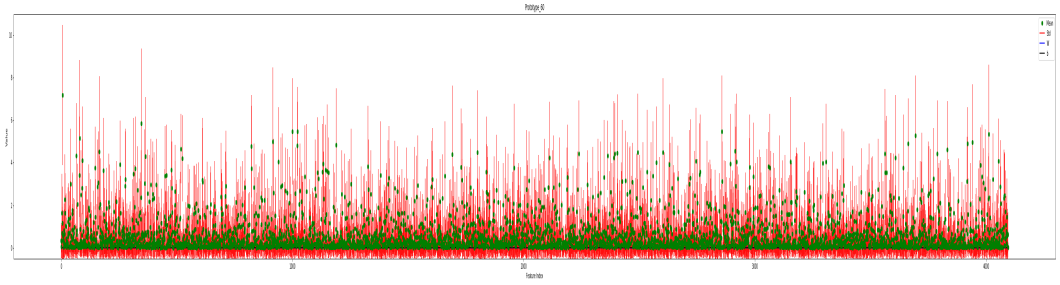
C.3 Protótipos no banco de dados ImageNet



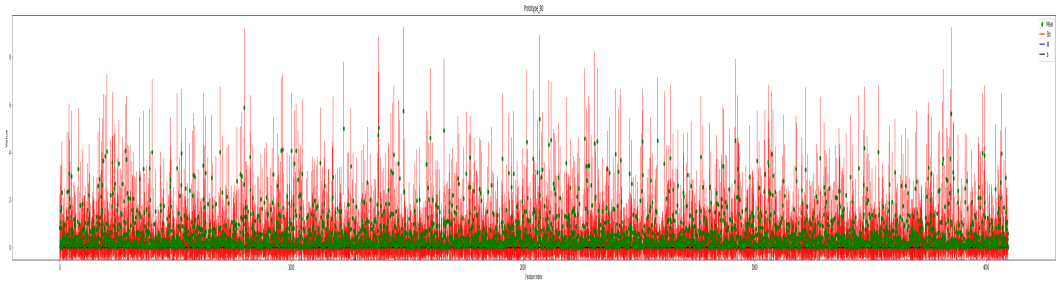
n02447366_15_badger



n02110341_40_dalmatian



n02132136_60_brown_bear



n02412080_80_ram

Figura C.5: Exemplos de protótipos calculados nas categorias c_{15}, c_{40}, c_{60} e c_{80} no banco de dados ImageNet usando o modelo VGG16. Em cada representação (de 4096 dimensões), mostram-se o id, índice e rótulo das categorias apresentadas.

Apêndice D

Comportamento prototípico

D.1 Banco de dados MNIST

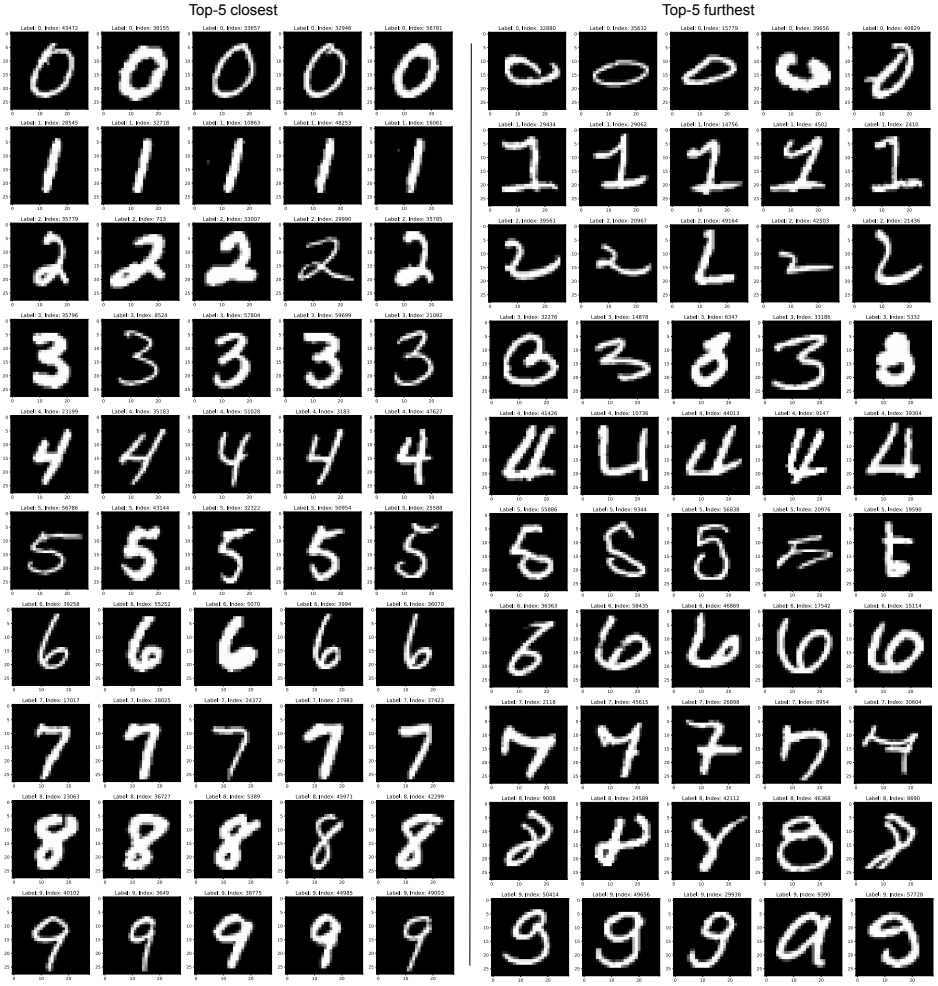


Figura D.1: *Exemplos do comportamento prototípico alcançado pelo modelo do protótipo proposto no banco de dados MNIST usando o modelo simples-MNIST. (Na esquerda)* mostra-se da esquerda para a direita o top-5 dos elementos mais próximos do *protótipo semântico* da categoria; *(Na direita)* os 5 elementos mais distantes do *protótipo semântico* da categoria.

D.2 Banco de dados CIFAR10

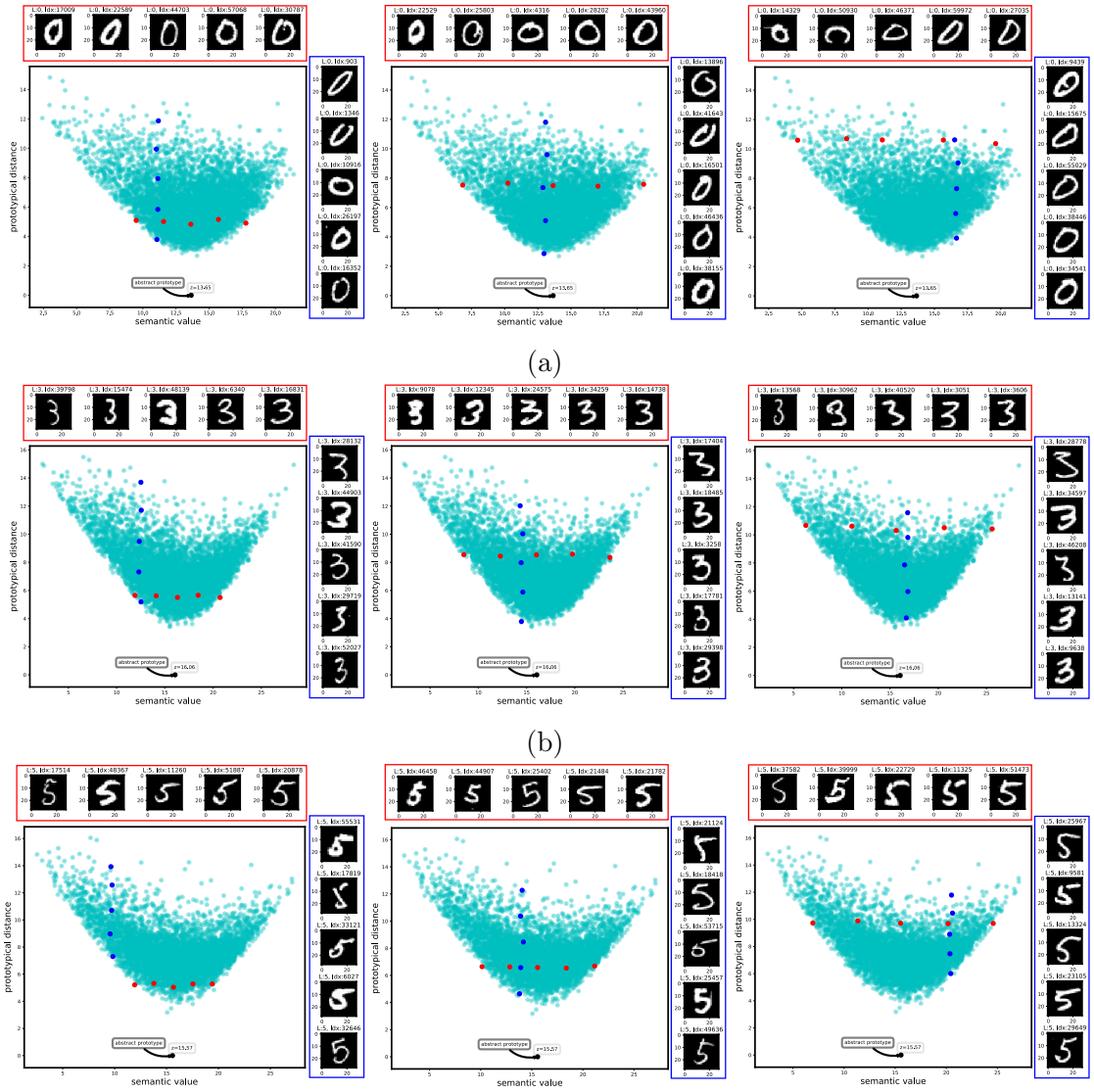
Figura D.2: *Exemplos do comportamento prototípico alcançado pelo modelo do protótipo proposto no banco de dados CIFAR10 usando o modelo simples-CIFAR10. (Na esquerda)* mostra-se da esquerda para a direita o top-5 dos elementos mais próximos do *protótipo semântico* da categoria; *(Na direita)* os 5 elementos mais distantes do *protótipo semântico* da categoria.

D.3 Banco de dados ImageNet

Figura D.3: *Exemplos do comportamento prototípico alcançado pelo modelo do protótipo proposto em uma mostra de categorias do banco de dados ImageNet usando o modelo VGG16. (Na esquerda)* mostra-se da esquerda para a direita o top-5 dos elementos mais próximos do *protótipo semântico* da categoria; *(Na direita)* os 5 elementos mais distantes do *protótipo semântico* da categoria.

Apêndice E

Análise das pontuações de tipicidade visual

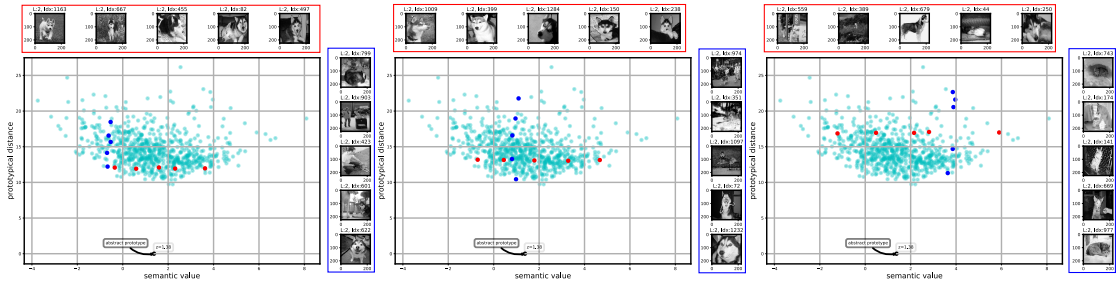


(a)

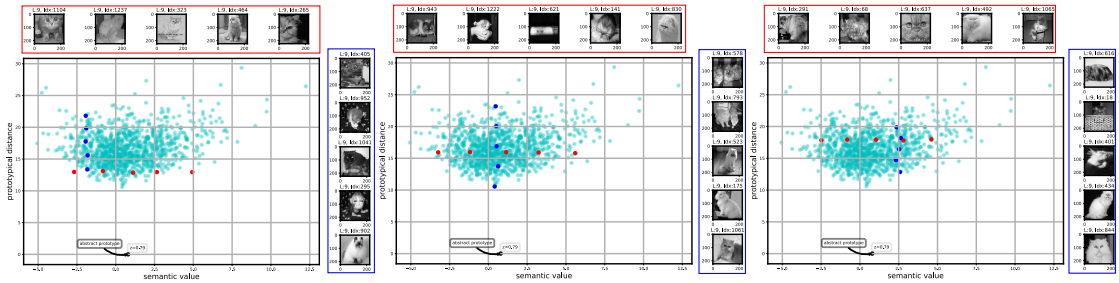
(b)

(c)

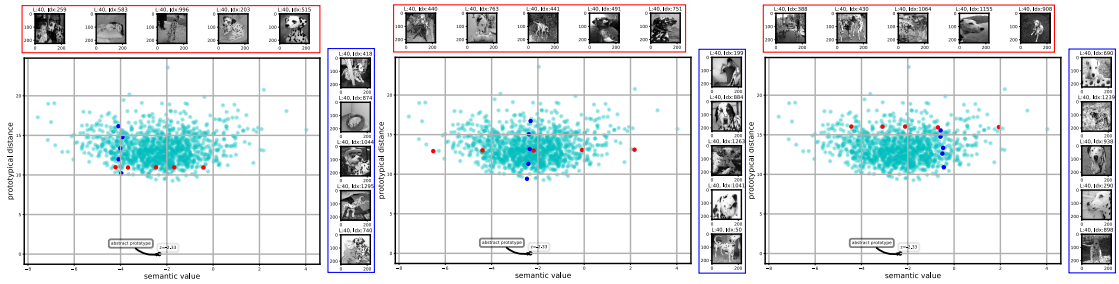
Figura E.1: Análise da tipicidade visual em algumas categorias do banco de dados MNIST usando o modelo *simples-MNIST*. Observe-se como imagens de objetos com a mesma distância prototípica (em vermelho) são visualmente semelhantes; e membros da categoria com distância prototípica diferente e valor semântico semelhante (em azul) são visualmente diferentes. Note-se também que a tipicidade visual da imagem diminui à medida que a distância prototípica aumenta.



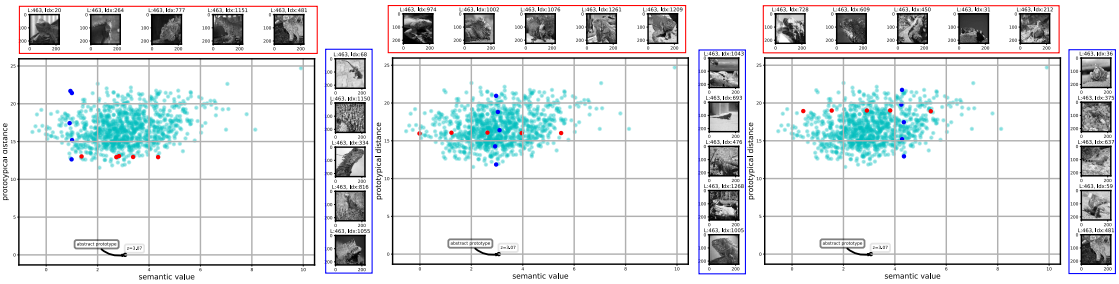
(a) Categoria 2 -*Siberian husky* (n02110185)



(b) Categoria 9 - *Persian cat* (n02123394)



(c) Categoria 40 - *Dalmatian* (n02110341)



(d) Categoria 463 - *Common iguana* (n01677366)

Figura E.2: Análise da tipicidade visual em algumas categorias do banco de dados ImageNet usando o modelo VGG16. Observe-se como imagens de objetos com a mesma distância prototípica (em vermelho) são visualmente semelhantes; e membros da categoria com distância prototípica diferente e valor semântico semelhante (em azul) são visualmente diferentes. Note-se também que a tipicidade visual da imagem diminui à medida que a distância prototípica aumenta.

Apêndice F

Detalhes da Transformação $f(x)$

O fluxo interno da função de transformação proposta possui várias etapas (Ver Figura 6.2) que podem ser detalhadas como:

Passo 1 *Redimensionar o vetor m-dimensional de entrada.* Seja a matriz quadrada unitária $\chi_{r\times r}$ com dimensões $(r \times r)$ definidas para cada modelo CNN usado. Redimensionar o vetor m-dimensional de entrada para uma matriz $\alpha_{p\times q}$ onde as dimensões p, q são múltiplos de r e cumprem que $m = p \cdot q$. A transformação realiza-se mantendo as conexões existentes da estrutura anterior. A nova geometria move o ponto de ativação (z) para o centro da matriz $\alpha_{p\times q}$ mantendo as conexões com as características, agora alocadas em novas posições dentro da matriz. Note-se que esta simples transformação não afeta o *valor semântico* $(\hat{z})$ porque ainda responde Ã Equação 5.4.

Passo 2 *Representação matricial.* Baseado na representação geométrica anterior é possível transformar facilmente a representação da Equação 5.4 na representação matricial da Equação F.1:

$$\vec{z_i} = \Omega_i \odot \alpha + \bar{b_i} \tag{F.1}$$

onde $\bar{b_i}$ constitui a matriz constante $\frac{b_i}{m}$, e $\odot$ representa *o produto Hadamard.*

A matriz semântica correspondente Ã categoria c_i $(\vec{z_i})$ constitui -exatamente- o *vetor semântico* $(\vec{z})$ redimensionado. A nova representação ainda preserva o valor semântico da categoria: $\hat{z} = \sum \vec{z}$. Aliás, a *matriz de direções* $(\Theta_{r\times r})$ se calcula baseada na posição de cada elemento da matriz quadrada unitária $\chi_{r\times r}$ com relação a seu centro. Como resultado final desta etapa, são obtidas quatro matrizes: α, Ω_i, $\bar{b_i}$, $\Theta_{r\times r}$.

Matriz de direções. A Figura F.1 mostra os detalhes do cálculo da matriz de direções $(\Theta_{r\times r})$. A matriz de direções contém os ângulos de cada característica com

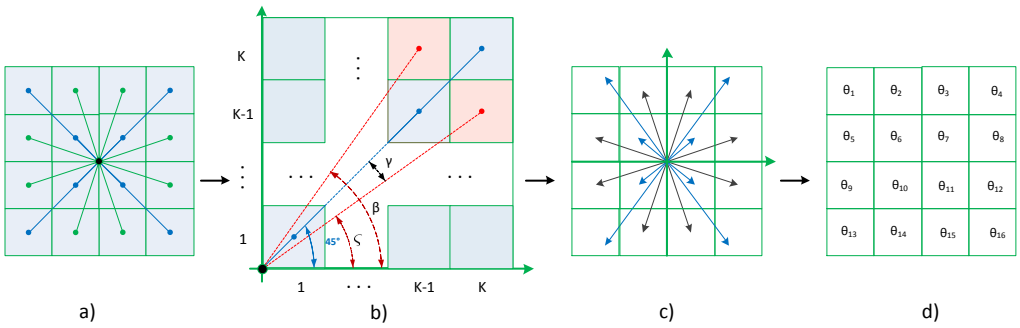


Figura F.1: *Detalhes da construção da matriz de direções* $\Theta_{r\times r}$. *a)* matriz quadrada $\chi_{r\times r}$; mostra-se em *azul* as conexões das posições localizadas nas diagonais da matriz quadrada, em *verde* o resto das outras conexões. *b)* cálculo de ângulos únicos para cada uma das conexões localizadas na diagonal da matriz (exemplo para 45°). *c)* vetores de direções resultantes para cada uma das posições da matriz quadrada $\chi_{r\times r}$ (vetores azuis para posições diagonais e vetores pretos para as posições que não estão na diagonal). *d)* matriz de direções $\Theta_{r\times r}$. Fonte: Elaborado pelo autor.

relação ao *eixo x* que define o centro da matriz quadrada unitária $\chi_{r\times r}$. Para as características com posições que não estão alocadas na diagonal da matriz unitária $\chi_{r\times r}$, são calculados os ângulos correspondentes (θ) formados pela conexão com o centro de $\chi_{r\times r}$ e o eixo de coordenada x (Ã s conexões verdes na Figura F.1 a) correspondem os vetores pretos na Figura F.1 c)). No caso das posições que se encontram na diagonal, constitui um pré-requisito encontrar *ângulos únicos* para cada uma desas características.

Seja $K = r/2$, e os ângulos $\varsigma = max(\{\theta \in \Theta_{r\times r} : 0 < \theta < 45°\})$ *e* $\beta = min(\{\theta \in \Theta_{r\times r} : 45° < \theta < 90°\})$ (Ver Figura F.1 b). Os ângulos correspondentes Ã s características localizadas na diagonal são uniformemente distribuídos entre os ângulos $[\varsigma, \beta]$ com um ângulo de $2\gamma/(K+1)$, onde $\gamma = (\beta - \varsigma)/2$. As características localizadas na diagonal entre as posições $[1, K/2]$ estão uniformemente distribuídas entre os ângulos $[\varsigma, 45°]$ e as características posicionadas entre $[K/2+1, K]$ são distribuídas entre os ângulos $[45°, \beta]$.

Passo 3 *Gradiente semântico.* Calcular o gradiente (g^{jk}) para cada matriz $\chi_{r\times r}$ mapeada nas matrizes α, Ω_i, $\bar{b_i}$ nas posições j, k. Semelhante ao algoritmo SIFT (Lowe, 2004) calcular a *magnitude, direção* e *orientação* de cada característica para todas as localizações da geometria (Ver Figura F.2). A direção do vetor da característica na posição x,y armazena-se em $\Theta_{r\times r}(x, y)$. A magnitude e a orientação estão baseadas no valor semântico da característica unitária $\vec{z}_i^{jk}(x, y)$. Os vetores projetados, uma vez que são gerados a partir de $\vec{z}_i$, possuem a propriedade de preservar, em conjunto, o valor semântico da categoria c_i: $\sum\sum_j\sum_k g^{jk} = \sum\sum_j\sum_k \vec{z}_i^{jk} = \sum \vec{z}_i = \hat{z}_i$.

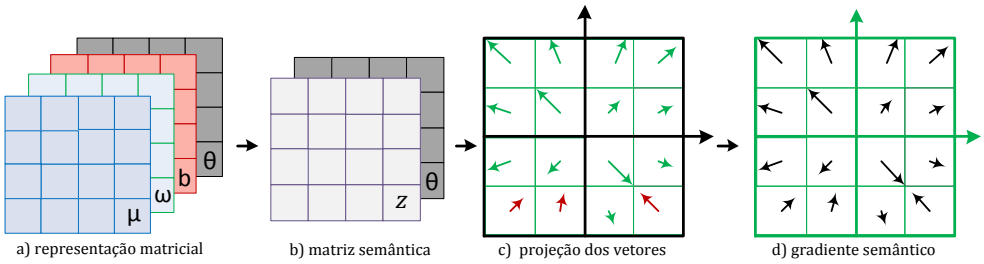


Figura F.2: *Cálculo do gradiente semântico.* Mostra-se o cálculo do *gradiente semântico* do exemplo apresentado na Figura 6.2. *a)* Conjunto de matrizes α^{jk}, Ω_i^{jk}, $\bar{b_i^{jk}}$ resultantes de mapear na posição j, k a matriz unitária $\chi_{r\times r}$ e $\Theta_{r\times r}$. *b)* cálculo da *matriz semântica* ($\vec{z}^{jk}$) construída a partir da Equação F.1. *c)* vetores projetados com as direções armazenadas em $\Theta_{r\times r}$, magnitude $|\vec{z}^{jk}|$ e orientação $sign(\vec{z}^{jk})$. Os vetores com orientações negativas são representados em *vermelho* e os vetores com orientações positivas são mostrados em *verde*. *d) gradiente semântico* resultante da projeção dos vetores calculados. Fonte: Elaborado pelo autor.

Passo 4 *Redução do gradiente semântico.* Reduzir o gradiente semântico para oito (8) vetores localizados nas orientações (θ) múltiplos de 45 graus. O conjunto dos oito vetores finais (v) calcula-se como a adição, no sentido anti-horário, dos vetores do gradiente com orientações entre os novos vetores projetados e seus vizinhos:

$$v_{\theta_l} = \sum_{\theta=\theta_l-45}^{\theta_l} g^{jk}(\theta) \tag{F.2}$$

$\forall\theta \in \Theta_{r\times r} : \theta_l - 45 < \theta \leq \theta_l$ onde $\theta_l = l \cdot \frac{\pi}{4}, \forall l = 1, ..., 8.$

Passo 5 *Assinatura do descritor.* Finalmente, construímos a assinatura unitária como o vetor de 8 dimensões (das magnitudes e das direções) projetado na orientação correspondente ($assinatura^{\,jk}(l) = |v_{\theta_l}| \cdot sign(v_{\theta_l})$ com $\theta_l = l \cdot \frac{\pi}{4}, \forall l = 1, ..., 8$). A assinatura do Descritor Semântico Global (ψ) correspondente ao vetor m-dimensional de entrada é construída pela concatenação das assinaturas unitárias ($assinatura^{\,jk}$) resultantes dos passos 3 e 4, para cada uma das matrizes mapeadas em α, Ω_i, $\bar{b_i}$ usando a matriz unitária ($\chi_{r\times r}$). O Algoritmo 3 resume o conjunto de passos da transformação $f(x)$ proposta.

Apêndice G

Relação entre as distâncias dos espaços métricos

Neste apêndice apresenta-se experimentos que exemplificam a relação existente entre os valores das distâncias dos espaços métricos mapeados usando as projetações $\rho : (F_{c_i}, \delta) \rightarrow (\mathbb{R}^2, L_1)$ e $\lambda : (\psi_{c_i}, L_1) \rightarrow (\mathbb{R}^2, L_1)$. Apresenta-se, para cada projeção no espaço métrico $(\mathbb{R}^2, L_1)$, exemplos das relações de distâncias correspondentes.

G.1 Mapeando o espaço das características CNN

Visualizar a estrutura interna de uma categoria específica resume-se a visualizar no espaço métrico $(\mathbb{R}^2, L_1)$ cada elemento da i-ésima categoria $c_i \in C$ mapeado através da função ρ. O anterior é consequência de que a *função continua* $\rho : (F_{c_i}, \delta) \rightarrow (\mathbb{R}^2, L_1)$ constitui um mapeamento do espaço métrico (F_{c_i}, δ) para o espaço métrico $(\mathbb{R}^2, L_1)$. Se $o_1, o_2 \in O_{c_i}$ e $p_1 = \rho(o_1)$, $p_2 = \rho(o_2)$; então a relação entre as métricas de distância δ e L_1 pode ser expressada como: $\delta(o_1, o_2) \leq L_1(p_1, p_2) \leq 2\delta(o_1, o_2)$ (Ver Proposição 1).

As Figuras G.1 e G.2 mostram exemplos da relação existente entre os valores das distâncias δ e L_1 para as categorias c_5 e c_{40} dos bancos de dados MNIST e ImageNet respectivamente. Apresenta-se, para cada membro com índice nos fragmentos das categorias mostradas, os valores da *distância prototípica* no domínio F_{c_i} e os valores da distância L_1 no domínio $\mathbb{R}^2$ com relação ao protótipo mapeado $\rho(P_i)$ da categoria. Mostram-se os valores da distância L_1 na cor laranja; e os valores limites inferior e superior da *distância prototípica* nas cores azul e verde respectivamente. Note-se como os valores das distâncias em ambos os espaços métricos cumprem com a relação apresentada na Proposição 1.

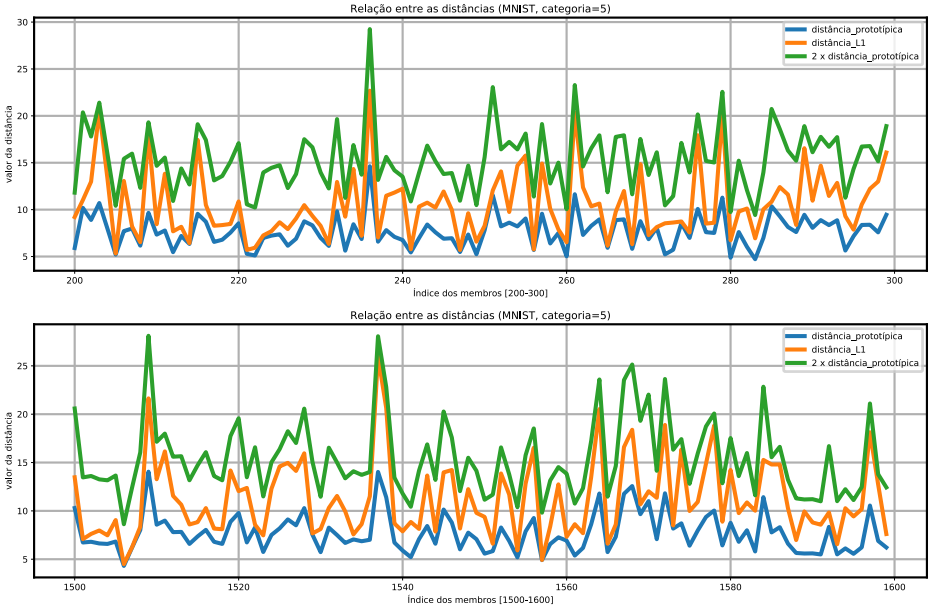


Figura G.1: *Exemplo da relação entre a distância prototípica e a distância L_1 para a categoria c_5 do banco de dados MNIST.* Mostram-se as distâncias com relação ao protótipo P_5 dos membros com índice nos fragmentos ([200-300],[1500-1600]) da categoria c_5.

Figura G.2: *Exemplo da relação entre a distância prototípica e a distância L_1 para a categoria c_{40} do banco de dados ImageNet usando o modelo VGG16.* Mostram-se as distâncias com relação ao protótipo P_{40} dos membros com índice nos fragmentos ([100-200],[600-700]) da categoria c_{40}.

G.2 Mapeando o domínio das assinaturas GSDP

Os experimentos nesta seção visam visualizar a estrutura semântica interna da categoria através do significado semântico encapsulado nas assinaturas de seus membros. Similar à abordagem usada no espaço métrico das características dos objetos (Ver Capítulo 5); visualizar a estrutura interna de uma categoria consiste na visualização do conteúdo semântico das assinaturas de cada elemento da categoria (assinaturas do espaço métrico (ψ_{c_i}, L_1)). Ou seja, é suficiente visualizar no espaço métrico $(\mathbb{R}^2, L_1)$ as assinaturas dos elementos da categoria $(c_i \in C)$ mapeadas através da função contínua $\lambda : (\psi_{c_i}, L_1) \rightarrow (\mathbb{R}^2, L_1)$ (Ver Proposição 2).

A função continua λ baseia-se nas relações existentes entre os valores da métrica L_1 nos espaços métricos (ψ_{c_i}, L_1) e $(\mathbb{R}^2, L_1)$. As Figuras G.3 e G.4 mostram exemplos da relação existente entre os valores da métrica L_1 para assinaturas das categorias c_5 e c_{40} dos bancos de dados MNIST e ImageNet respectivamente. Apresentam-se, para cada membro com índice nos fragmentos das categorias mostradas, os valores correspondentes da métrica de distância L_1 no domínio ψ_{c_i} e no domínio $\mathbb{R}^2$. No domínio ψ_{c_i} os valores da métrica L_1 se calculam entre as assinaturas dos membros da categoria (ψ_{c_i}) e a assinatura do protótipo correspondente (ψ_{P_i}). No domínio $\mathbb{R}^2$ os valores

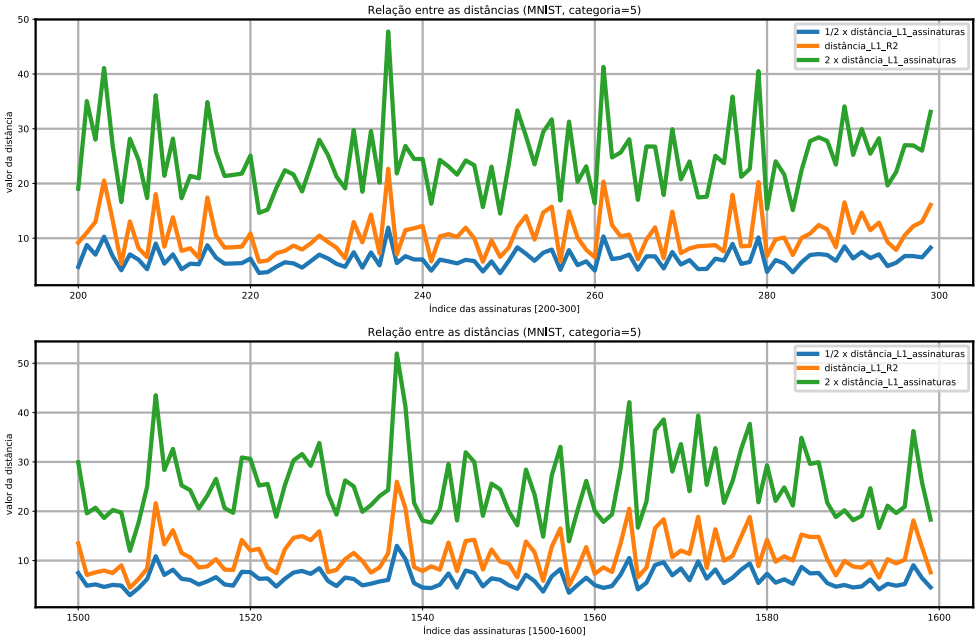


Figura G.3: *Exemplo da relação entre as distâncias L_1 dos domínios ψ_{c_5} e $\mathbb{R}^2$ para a categoria c_5 do banco de dados MNIST.* Mostram-se as distâncias, com relação à assinatura do protótipo abstrato P_5, dos membros da categoria c_5 fragmentos ([200-300],[1500-1600]).

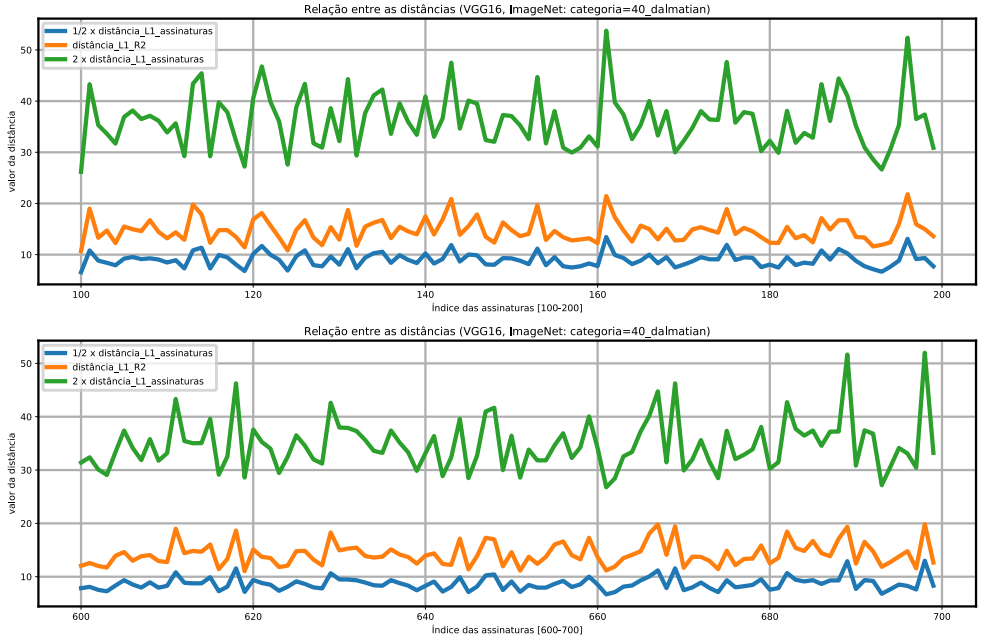


Figura G.4: *Exemplo da relação entre as distâncias L_1 dos domínios $\psi_{c_{40}}$ e $\mathbb{R}^2$ para a categoria c_{40} do banco de dados ImageNet usando o modelo VGG16.* Mostram-se as distâncias –com relação à assinatura do protótipo P_{40}– dos membros dos fragmentos da categoria c_{40} ([100-200],[600-700]).

da métrica L_1 se calculam com relação à assinatura do protótipo da categoria mapeada através da função $\lambda(\psi_{P_i})$. Nessas figuras mostram-se os valores da distância L_1 no domínio $\mathbb{R}^2$ na cor laranja, e os valores limites inferior e superior da distância L_1 no domínio ψ_{c_i} nas cores azul e verde respectivamente.

Apêndice H

Exemplos de organização prototípica da categoria

H.1 Categorias do banco de dados MNIST

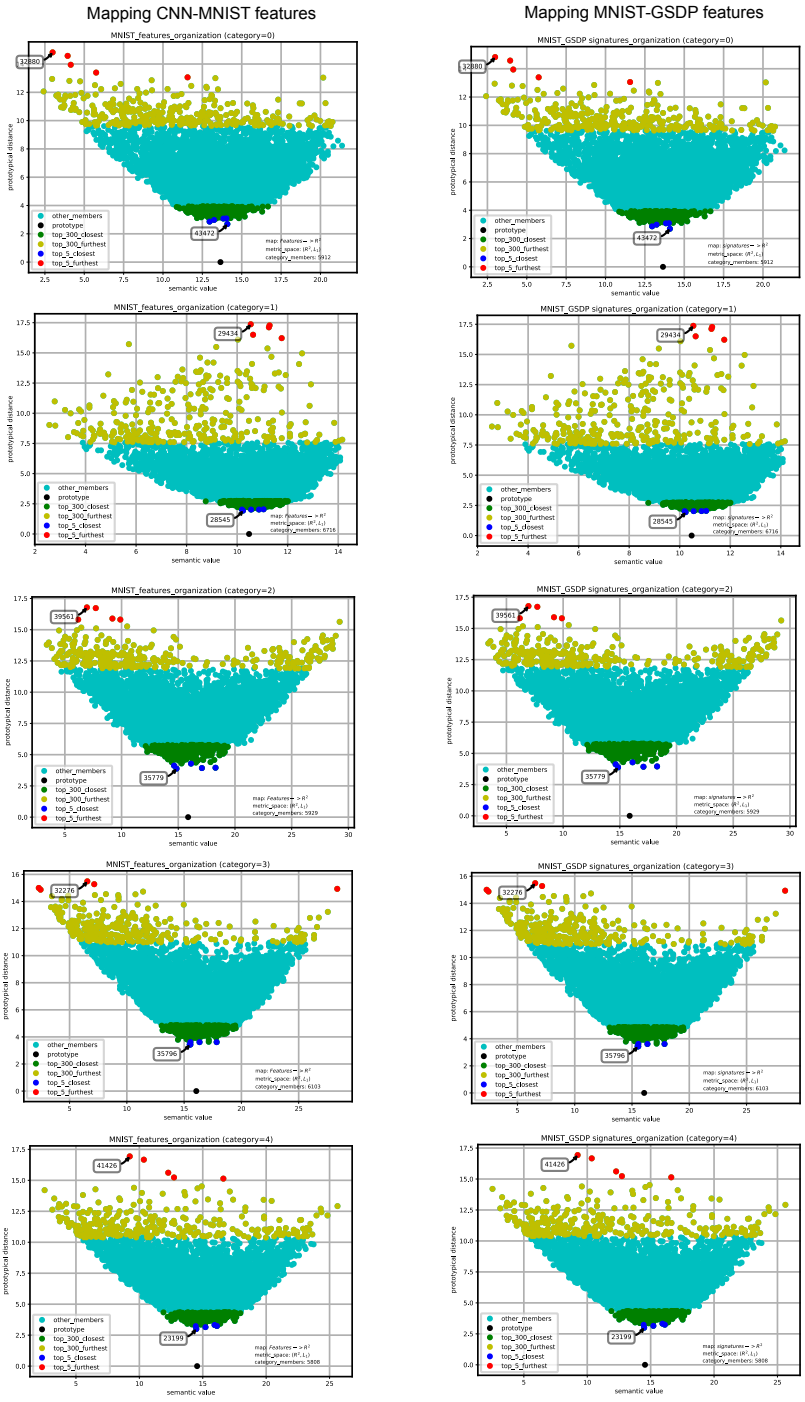


Figura H.1: Organização prototípica observada nas primeiras cinco categorias do conjunto de dados MNIST. Observe como a organização interna das categorias mapeadas no espaço métrico $(\mathbb{R}^2, L_1)$ a partir de assinaturas do descritor GSDP (na direita), é idêntica à organização interna das características CNN-MNIST mapeadas no espaço métrico $(\mathbb{R}^2, L_1)$ (na esquerda).

H.2 Categorias do banco de dados ImageNet

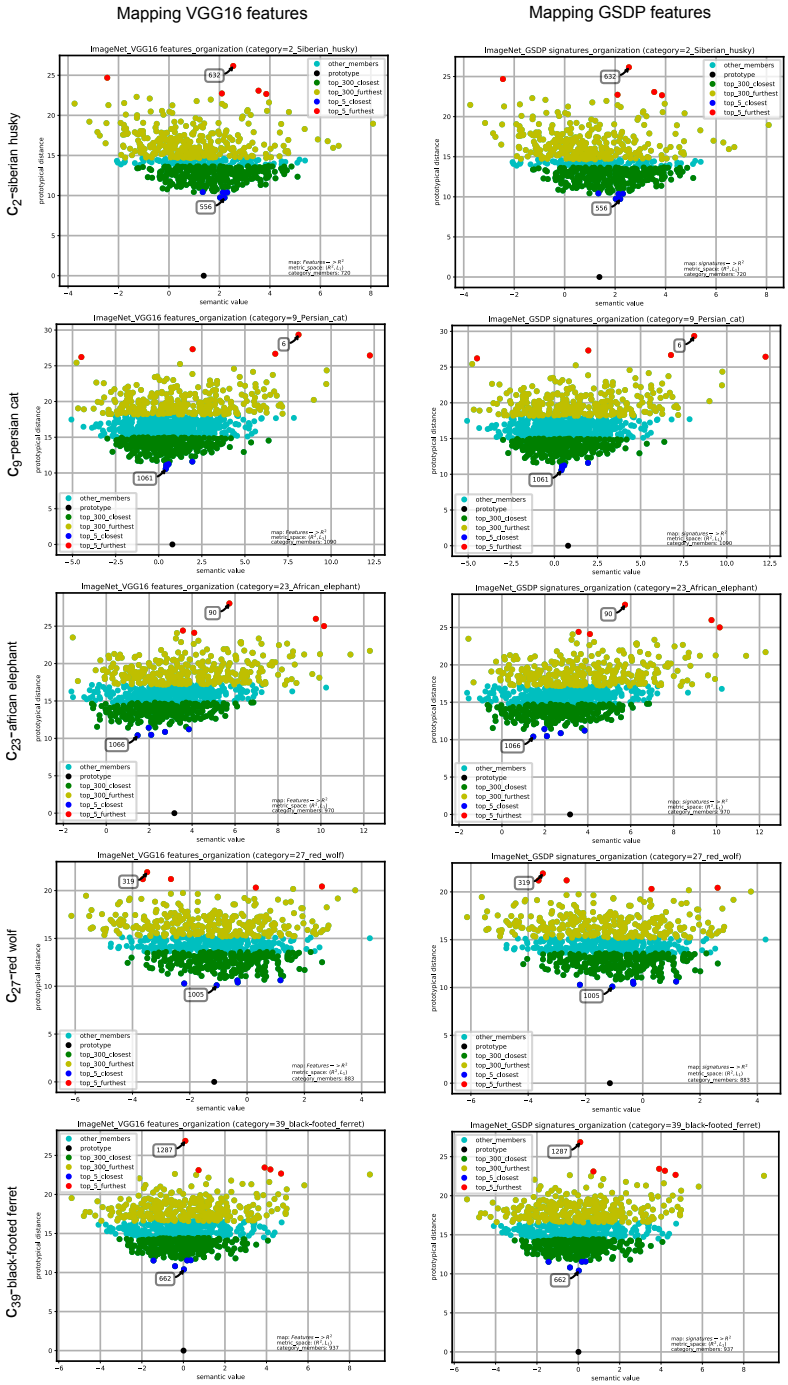


Figura H.2: Organização prototípica observada em uma amostra das categorias do conjunto de dados ImageNet. Observe como a organização interna das categorias mapeadas no espaço métrico $(\mathbb{R}^2, L_1)$ a partir de assinaturas do descritor GSDP (na direita), é idêntica à organização interna das características VGG16 mapeadas no espaço métrico $(\mathbb{R}^2, L_1)$ (na esquerda).

Apêndice I

Visualização t-SNE

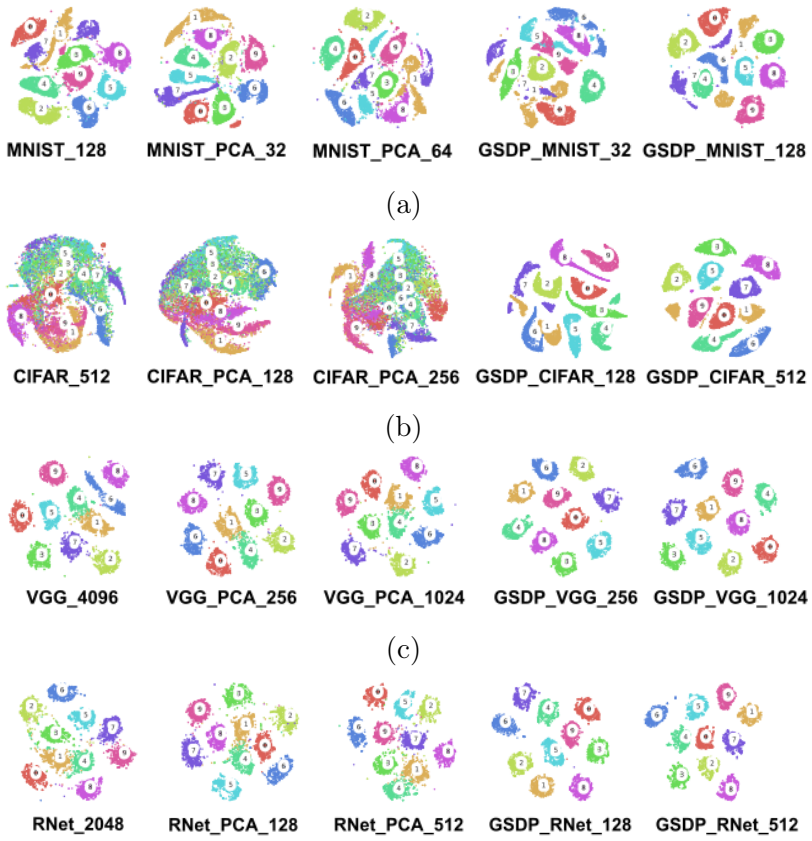


Figura I.1: *Visualização t-SNE. a)* visualização t-SNE de características construídas com o modelo *simples-MNIST* no conjunto de dados MNIST; *b)* visualização t-SNE de características construídas com o modelo *simples-CIFAR* no conjunto de dados CIFAR; *c,d)* visualização t-SNE das 10 primeiras categorias do conjunto de dados ImageNet usando características construídas com os modelos *VGG16* e *ResNet50*, respectivamente. O comprimento de cada característica mostra-se na legenda correspondente.

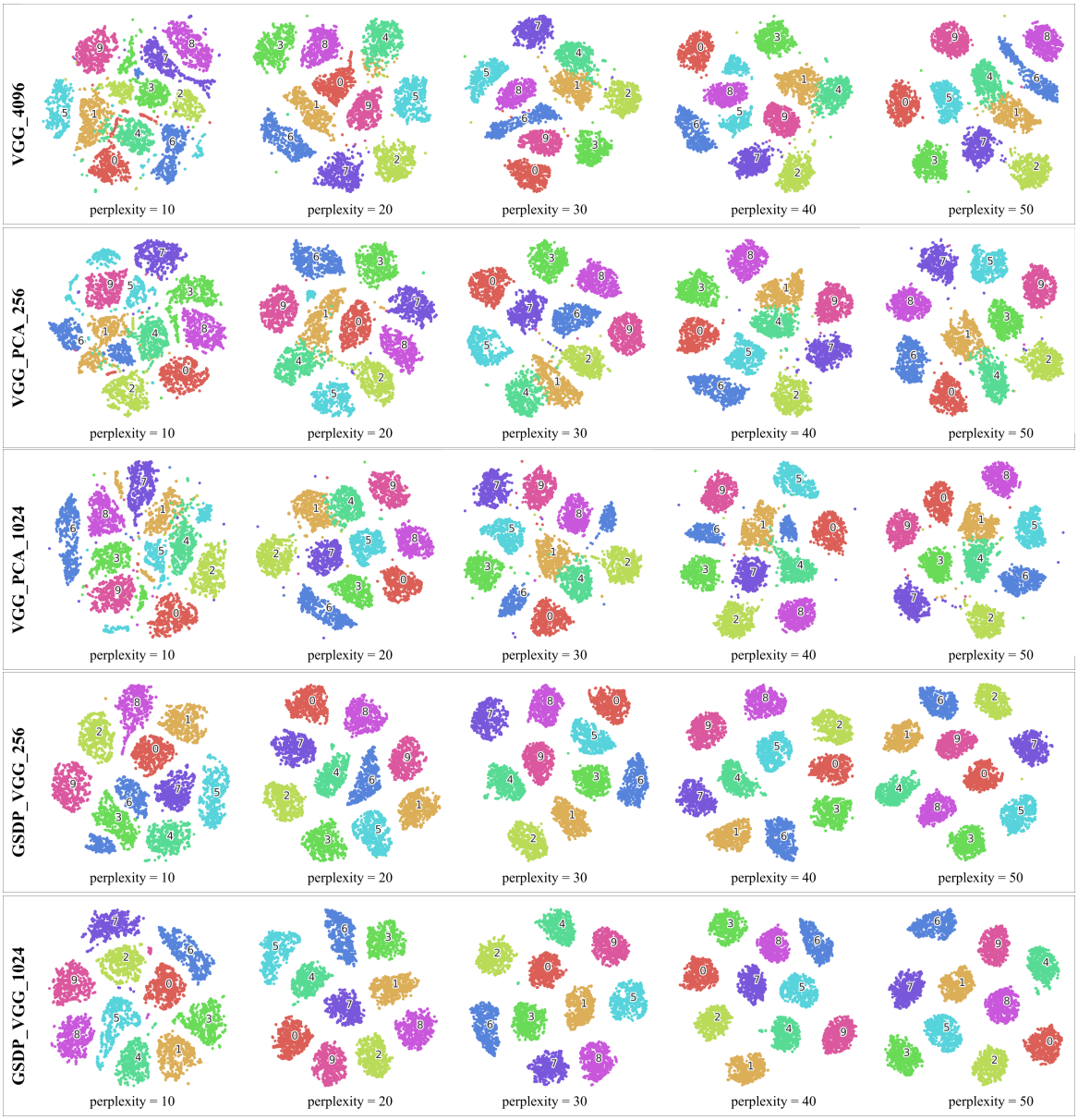


Figura I.2: Visualização t-SNE com a família de características do modelo VGG16 para as 10 primeiras categorias do conjunto de dados ImageNet e usando a distância euclidiana como medida de similaridade das características. Cada linha mostra a visualização t-SNE de cada tipo de característica VGG16 avaliada enquanto o valor de perplexidade aumenta.

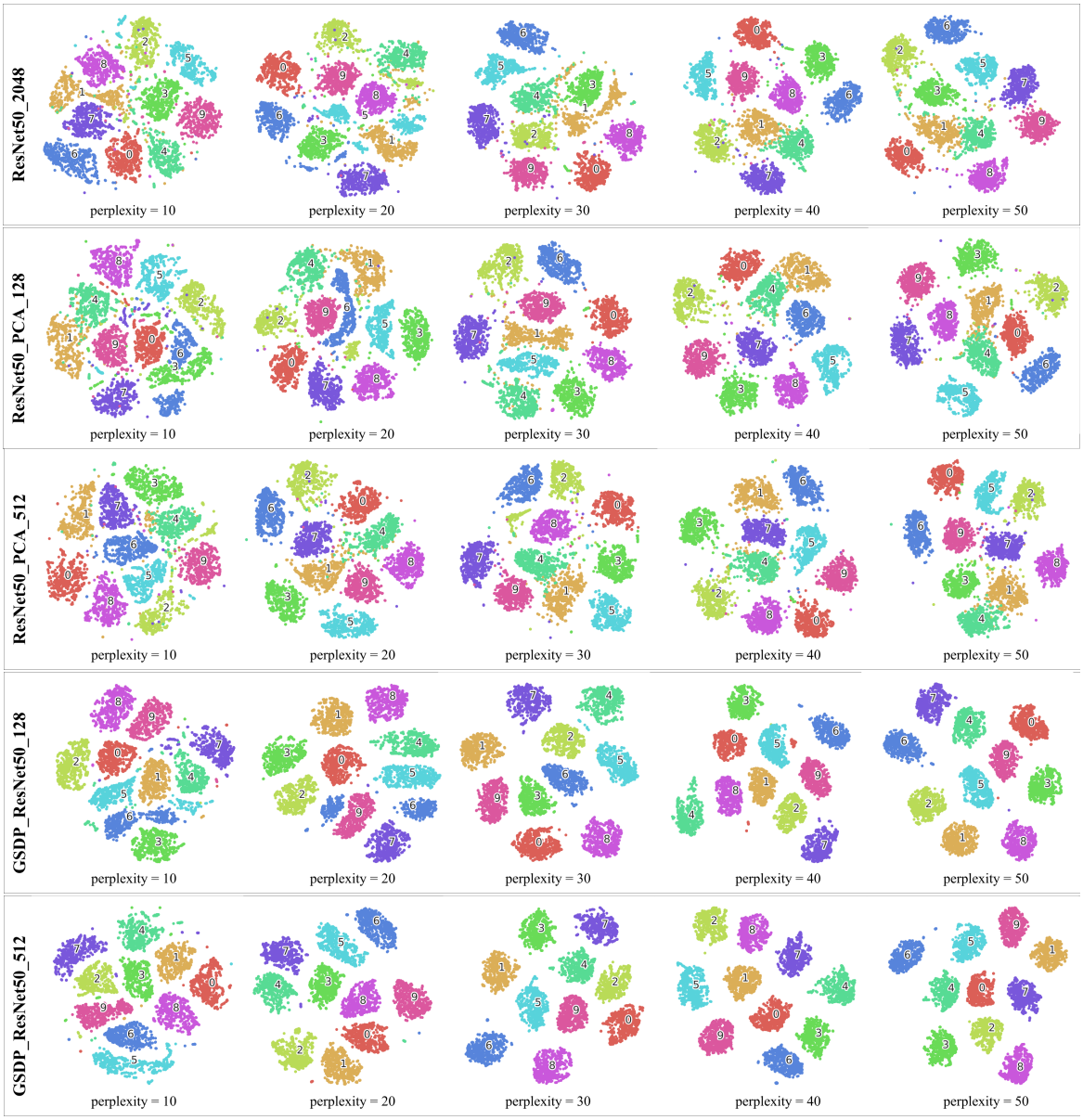
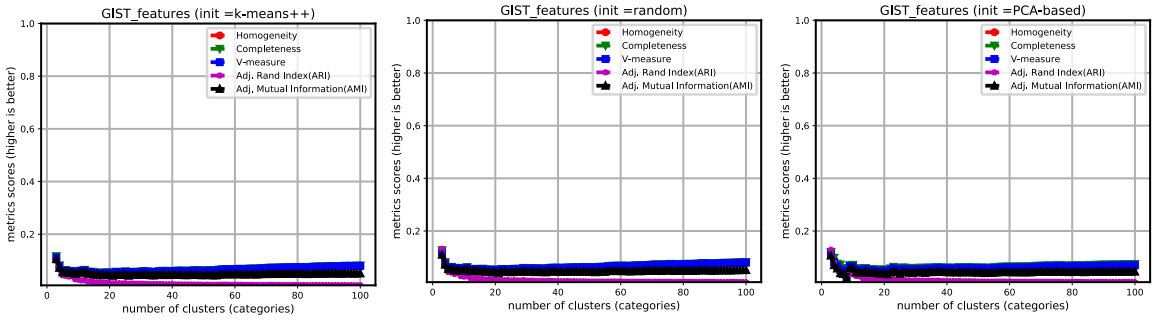


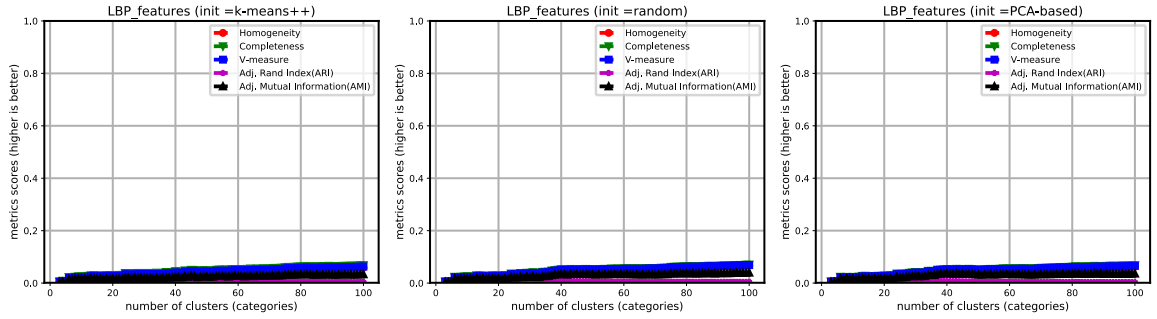
Figura I.3: Visualização t-SNE com a família de características do modelo ResNet50 para as 10 primeiras categorias do conjunto de dados ImageNet e usando a distância euclidiana como medida de similaridade das características. Cada linha mostra a visualização t-SNE de cada tipo de característica ResNet50 avaliada enquanto o valor de perplexidade aumenta.

Apêndice J

Avaliação do descritor em tarefas de agrupamento e classificação

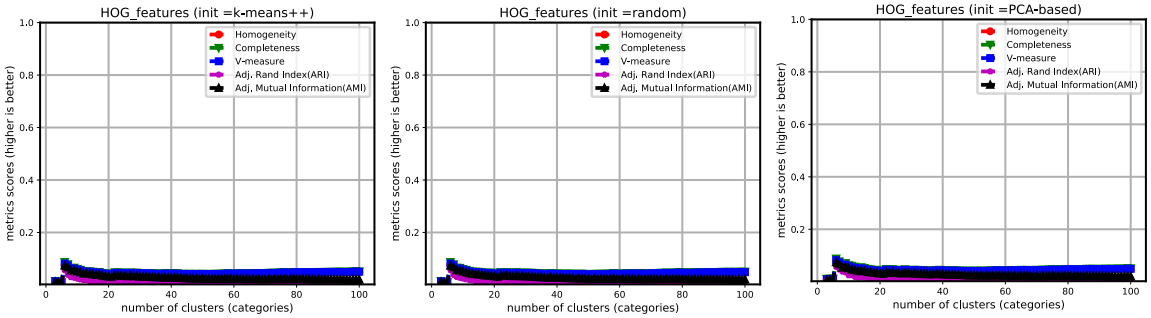


(a)

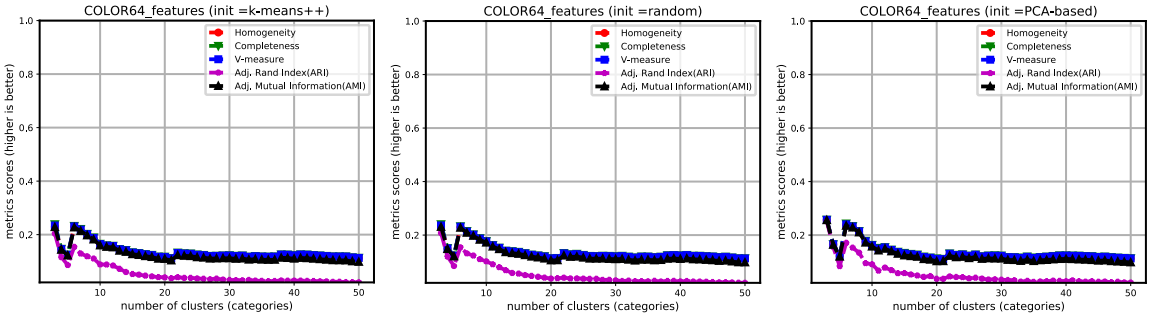


(b)

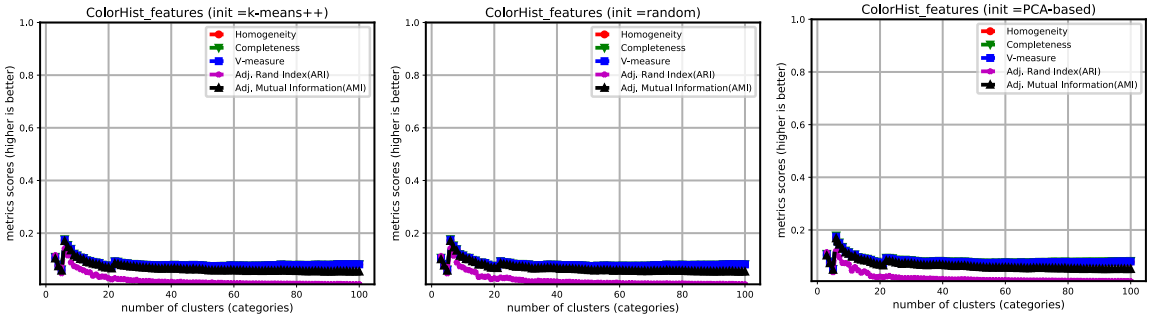
Figura J.1: Comportamento das métricas de agrupamento K-Means para várias assinaturas de descritores, quando o número de partições (categorias) aumenta em cada iteração do experimento (100). Os experimentos foram realizados usando várias estratégias de inicialização do algoritmo K-Means (*k-means++*, *random* e *pca-based*) para cada uma das representações de características: (a) GIST; (b) LBP.



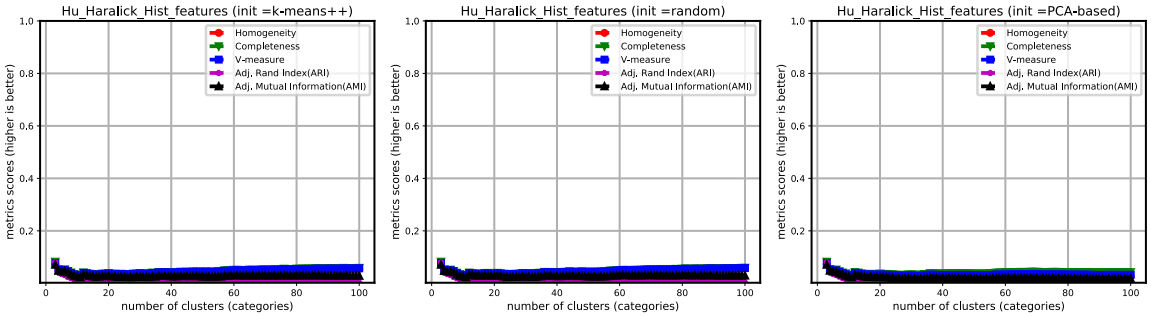
(a)



(b)

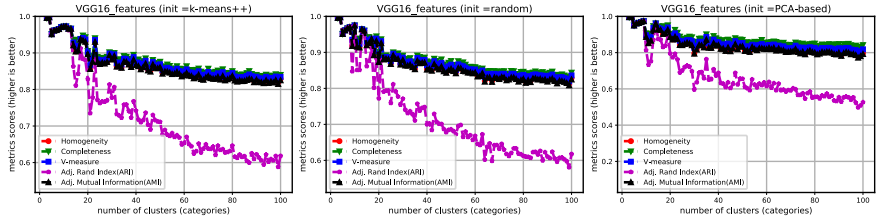


(c)

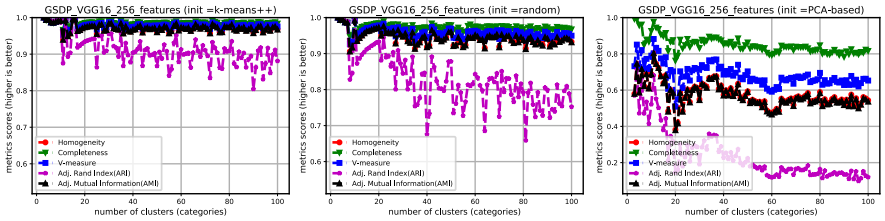


(d)

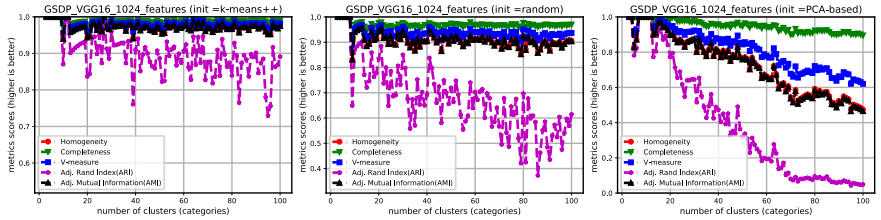
Figura J.2: Comportamento das métricas de agrupamento K-Means para várias assinaturas de descritores, quando o número de partições (categorias) aumenta em cada iteração do experimento (100). Os experimentos foram realizados usando várias estratégias de inicialização do algoritmo K-Means (*k-means++*, *random* e *pca-based*) para cada uma das representações de características: (a) HOG; (b) COLOR64; (c) ColorHist; (d) Hu-Haralick-Hist.



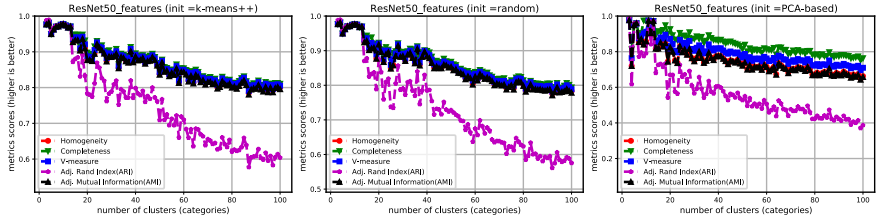
(a)



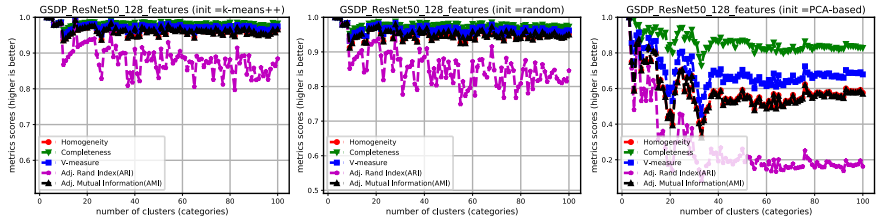
(b)



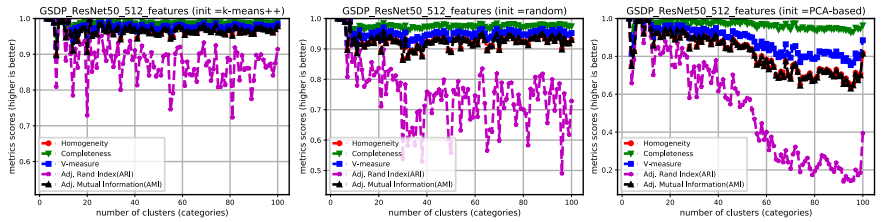
(c)



(d)



(e)



(f)

Figura J.3: Comportamento das métricas de agrupamento K-Means para várias assinaturas de descritores, quando o número de partições (categorias) aumenta em cada iteração do experimento (100). Os experimentos foram realizados usando várias estratégias de inicialização do algoritmo K-Means (*k-means++*, *random* e *pca-based*) para cada uma das representações de características: (a) VGG16; (b) GSDP_VGG_256; (c) GSDP_VGG_1024; (d) ResNet50; (e) GSDP_ResNet_128; (f) GSDP_ResNet_512.

Apêndice K

Histórico de treinamento dos modelos PS-Layer construídos

K.1 Modelo simples-MNIST e versões PS-Layer

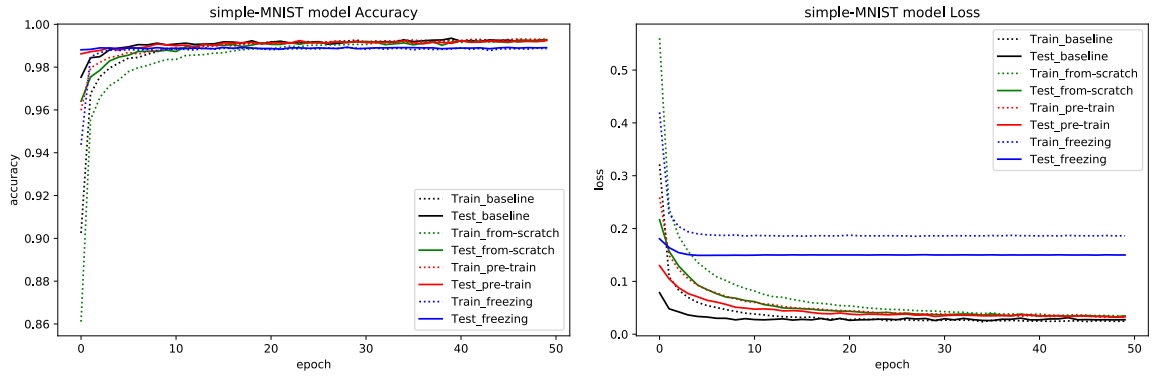


Figura K.1: Curvas do histórico de treinamento do modelo simples-MNIST e as versões PS-Layer que usam a distância prototípica. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados MNIST, respetivamente.

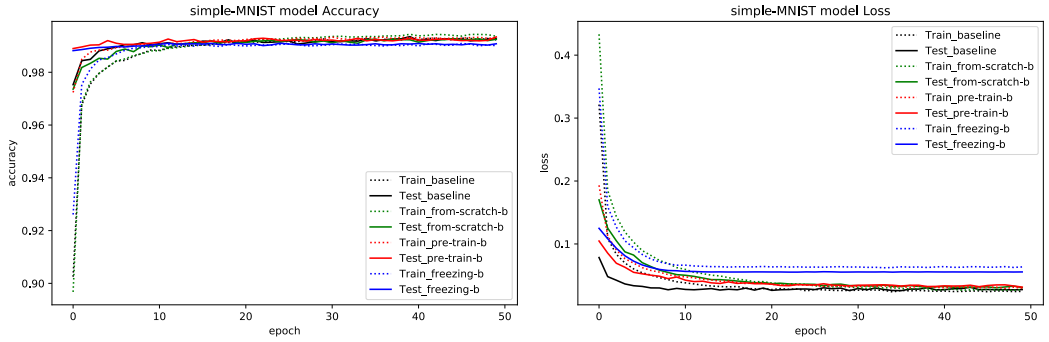


Figura K.2: Curvas do histórico de treinamento do modelo simples-MNIST e as versões PS-Layer que usam a distância prototípica penalizada. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados MNIST, respetivamente.

K.2 Modelo simples-CIFAR10 e versões PS-Layer

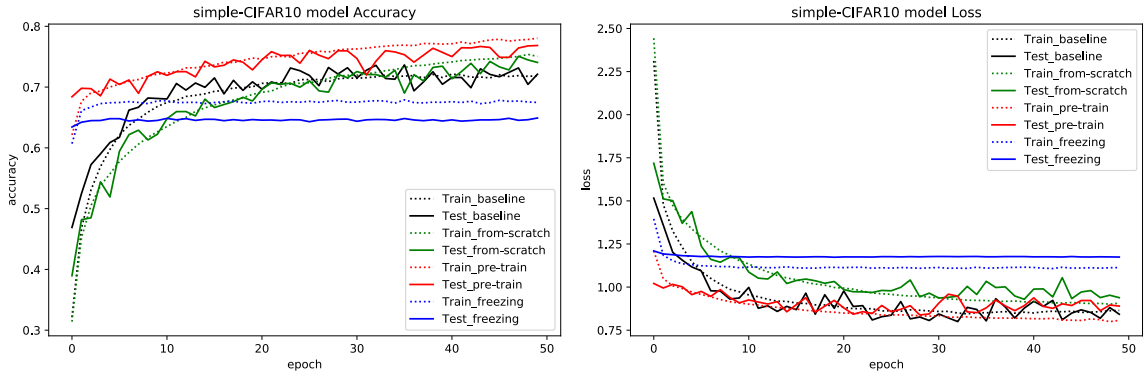


Figura K.3: Curvas do histórico de treinamento do modelo simples-CIFAR10 e as versões PS-Layer que usam a distância prototípica. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados CIFAR10, respetivamente.

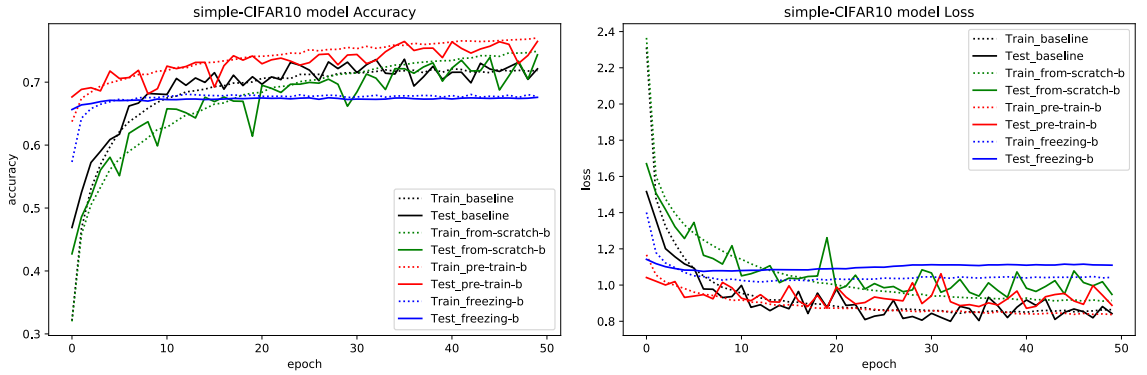


Figura K.4: Curvas do histórico de treinamento do modelo simples-CIFAR10 e as versões PS-Layer que usam a distância prototípica penalizada. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados CIFAR10, respetivamente.

K.3 Modelo VGG-CIFAR10 e versões PS-Layer

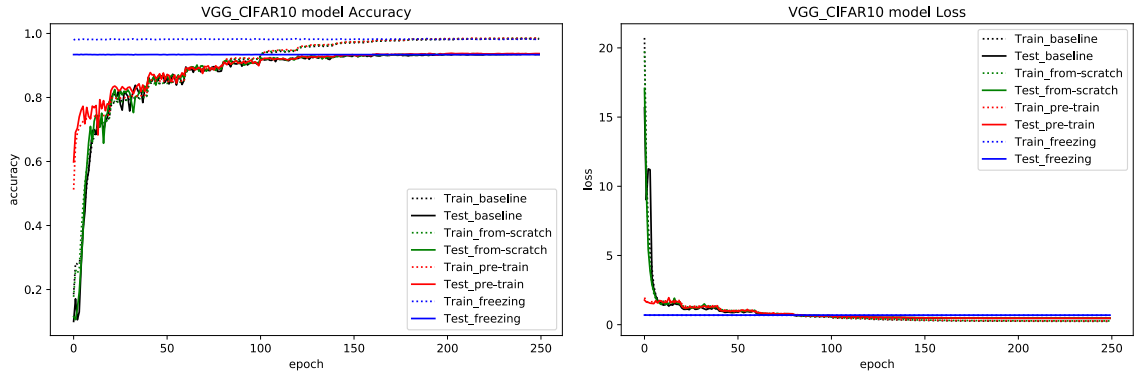


Figura K.5: Curvas do histórico de treinamento do modelo VGG-CIFAR10 e as versões PS-Layer que usam a distância prototípica. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados CIFAR10, respetivamente.

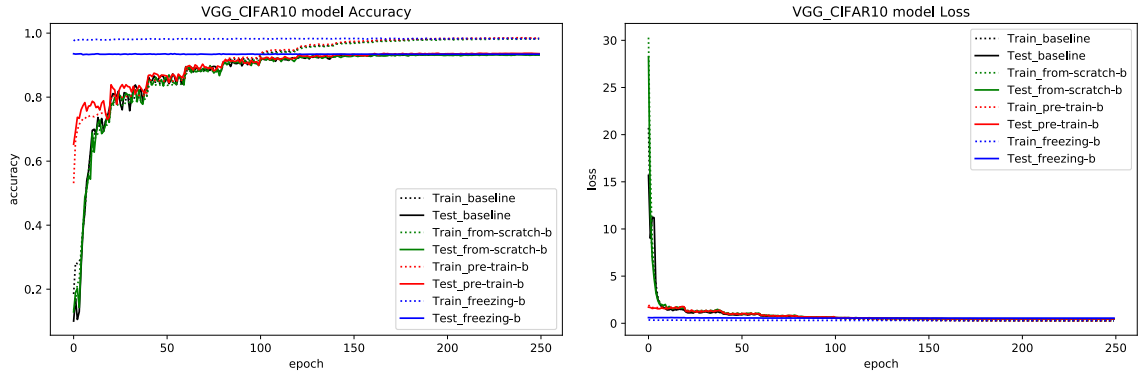


Figura K.6: Curvas do histórico de treinamento do modelo VGG-CIFAR10 e as versões PS-Layer que usam a distância prototípica penalizada. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados CIFAR10, respetivamente.

K.4 Modelo VGG-CIFAR100 e versões PS-Layer

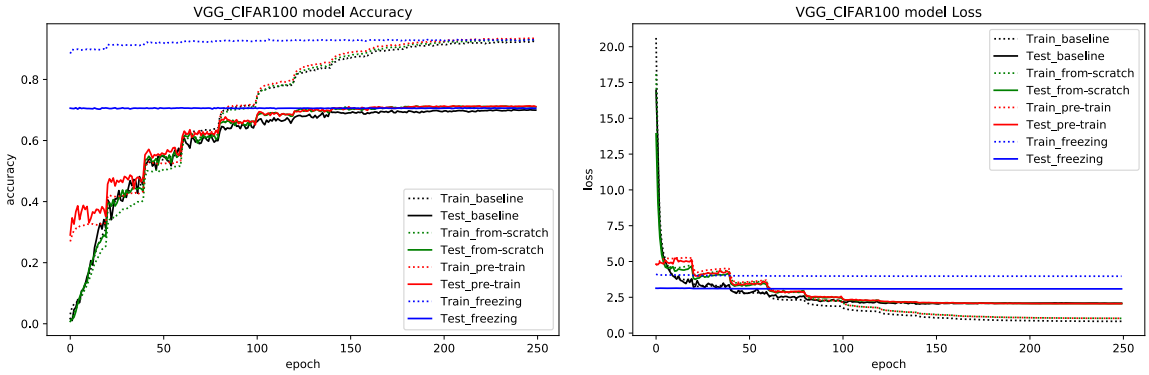


Figura K.7: Curvas do histórico de treinamento do modelo VGG-CIFAR100 e as versões PS-Layer que usam a distância prototípica. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados CIFAR100, respetivamente.

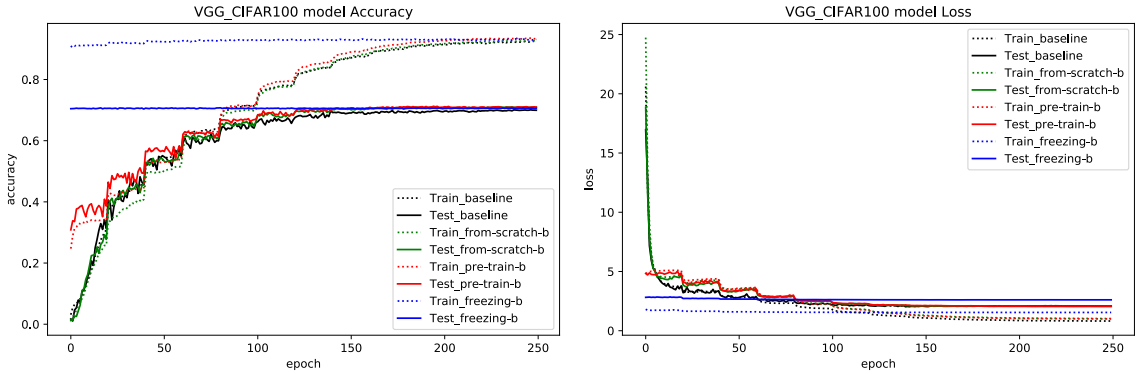


Figura K.8: Curvas do histórico de treinamento do modelo VGG-CIFAR100 e as versões PS-Layer que usam a distância prototípica penalizada. *Train* e *Test* representam o desempenho de cada versão de modelo treinada nos conjuntos de dados Treinamento e Validação do conjunto de dados CIFAR100, respetivamente.

Anexo A

Atributos métricos de uma boa característica

Good Feature Metric Attributes	Details
Scale invariance	Should be able to find the feature at different scales
Perspective invariance	Should be able to find the feature from different perspectives in the field of view
Rotational invariance	The feature should be recognized in various rotations within the image plane
Translation invariance	The feature should be recognized in various positions in the FOV
Reflection invariance	The feature should be recognized as a mirror image of itself
Affine invariance	The feature should be recognized under affine transforms
Noise invariance	The feature should be detectable in the presence of noise
Illumination invariance	The feature should be recognizable in various lighting conditions including changes in brightness and contrast
Compute efficiency	The feature descriptor should be efficient to compute and match
Distinctiveness	The feature should be distinct and detectable, with a low probability of mis-match, amenable to matching from a database of features
Compact to describe	The feature should not require large amounts of memory to hold details
Occlusion robustness	The feature or set of features can be described and detected when parts of the feature or feature set are occluded
Focus or blur robustness	The feature or set of features can be detected at varying degrees of focus (i.e, image pyramids can provide some of this capability)
Clutter and outlier robustness	The feature or set of features can be detected in the presence of outlier features and clutter

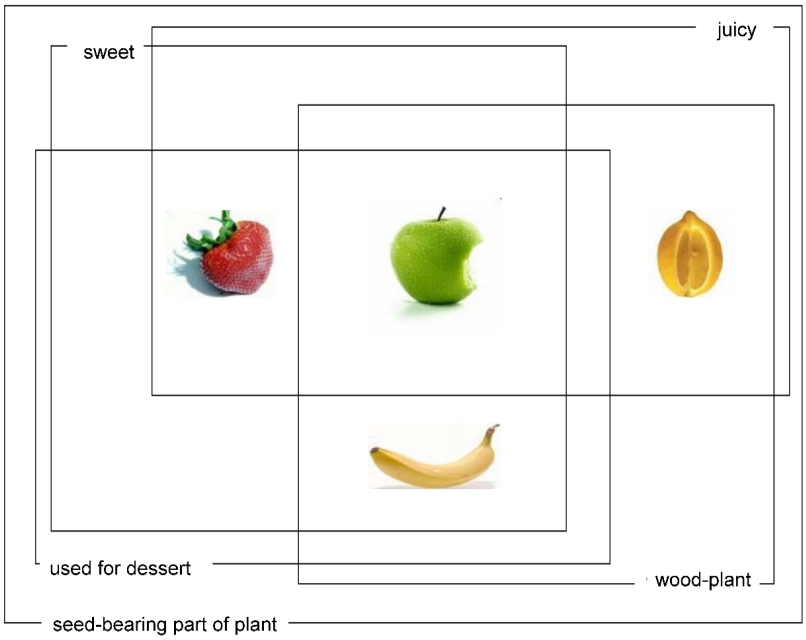
Figura A.1: Atributos métricos de uma boa característica. Fonte: Krig, 2014, p. 135.

Anexo B

Exemplo dos efeitos prototípicos

	edible seed-bearing part	of wood-plant	juicy	sweet	used as dessert
apple	+	+	+	+	+
strawberry	+	-	+	+	+
banana	+	+	-	+	+
lemon	+	+	+	-	-

(a)



(b)

Figura B.1: Efeitos prototípicos na categoria fruta. (a) Caracterização de membros da categoria fruta; (b) efeitos prototípicos na estrutura semântica interna da categoria fruta. Fonte: Geeraerts, 2010, p. 305.