UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Matheus Alves Diniz

Representação de Atributos Faciais nas Camadas, Canais e Neurônios de Redes Neurais de Reconhecimento Facial

Belo Horizonte
2020-06

Matheus Alves Diniz

Representação de Atributos Faciais nas Camadas, Canais e Neurônios de Redes Neurais de Reconhecimento Facial

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: William Robson Schwartz

Belo Horizonte
2020-06

Face attribute representation across the layers, channels and neurons of face recognition neural networks

Diniz, Matheus Alves.

D585f Face attribute representation across the layers, channels and neurons of face recognition neural networks [manuscrito] / Matheus Alves Diniz. – 2021.
xviii, 48 f. il.

Orientador:. William Robson Schwartz.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f.43-48
.
1. Computação – Teses. 2. Percepção visual – Teses. 3. Reconhecimento de faces – Teses. 4. Visão computacional – Teses. I. Schwartz, William Robson. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. III.Título.

CDU 519.6*82.10(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

FACE ATTRIBUTE REPRESENTATION ACROSS THE LAYERS, CHANNELS AND NEURONS OF FACE RECOGNITION NEURAL NETWORKS

MATHEUS ALVES DINIZ

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WILLIAM ROBSON SCHWARTZ - Orientador
Departamento de Ciência da Computação - UFMG

PROF. DAVID MENOTTI GOMES
Departamento de Informática - UFPR

PROF. ADRIANO ALONSO VELOSO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 31 de Março de 2021.

Resumo

As representações aprendidas por redes profundas são os descritores estado-da-arte para métodos de reconhecimento facial. Essas representações codificam características latentes que são difíceis de serem explicadas, o que compromete a confiança e interpretabilidade de suas predições. A maior parte das tentativas de se explicar essas características são técnicas de visualização, cuja principal limitação é relativa à sua subjetividade. Ao invés das visualizações, este trabalho propõe a utilização de camadas intermediárias da rede para classificar atributos faciais. A performance obtida por esses classificadores é utilizada como um indicador do quão bem aquele atributo é aprendido implicitamente naquela camada. Essa análise pode ainda ser combinada com uma técnica de seleção de variáveis para estabelecer precisamente a localização dos neurônios relevantes para cada atributo. De acordo com os experimentos, atributos que codificam gênero, utilização de óculos e chapéu podem ser preditos com uma acurácia superior a 96% através da saída de um único neurônio. Essa performance é apenas 3 pontos percentuais inferior a métodos estado da arte que foram supervisionados para predizer esses atributos, o que indica que estes atributos são muito bem definidos dentro da rede de reconhecimento facial.

Palavras-chave: atributos faciais, reconhecimento de faces, aprendizado profundo, entedimento de redes neurais

Palavras-chave: Atributos faciais, Reconhecimento facial, Aprendizado profundo, Entedimento de redes neurais

Abstract

Deeply learned representations are the state-of-the-art descriptors for face recognition methods. These representations encode latent features that are difficult to explain, compromising the confidence and interpretability of their predictions. Most attempts to explain deep features are visualization techniques that are often open to interpretation. Instead of relying only on visualizations, we use the outputs of hidden layers to predict face attributes. The obtained performance is an indicator of how well the attribute is implicitly learned in that layer of the network. Using a variable selection technique, we also analyze how these semantic concepts are distributed inside each layer, establishing the precise location of relevant neurons for each attribute. According to our experiments, gender, eyeglasses and hat usage can be predicted with over 96% accuracy even when only a single neural output is used to predict each attribute. This performance is less than 3 percentage points lower than the one achieved by deep supervised face attribute networks, which indicates that there exists neurons inside face recognition DCNNs encoding face attributes almost as accurately as DCNNs optimized specifically for these attributes.

Keywords: face attributes, face recognition, deep learning, neural networks understanding

Keywords: Face Recognition, Deep-Learning, Explainability

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Recent advances in deep learning provided more powerful and discriminative representations for modern face recognition systems [58; 5]. These representations, also known as deeply-learned features, are extracted from the hidden layers of Convolutional Neural Networks (CNNs), which are optimized to recognize thousands of individuals. Some training objective functions [41; 5; 6] can be used to precisely define what each neuron in the final layer of the network encodes. Due to the complexity of this objective function, intermediate layers are necessary to distill useful latent representations from the input [1]. Neurons from these layers, however, are not directly supervised, and consequently, these intermediate representations cannot be easily defined [56]. Thus, the usage of deeply-learned features in face recognition systems presents one important drawback: users have an insufficient understanding of the reasoning behind the decisions made by these systems, limiting both their acceptance [4] and improvement [20].

This thesis investigates the representation of several face attributes across the layers of deep face recognition networks. This investigation helps understanding what type of information is employed during recognition and how it impacts the deployment of these systems in scenarios with stricter restrisction such as in surveillance, where there is no guarantee of user cooperation.

1.1 Motivation and Scope

Most recent attempts to understand CNNs have focused on visualization techniques [44; 24; 25; 59]. The popularity of these techniques might be attributed to their flexibility, which allows the generation of an image for any neural pattern of the network such as the outputs of neurons, filters and layers. Each such generated image requires the interpretation of a trained user to extract useful insights from the synthetic visualizations. Consequently, these approaches are usually limited to a few qualitative examples, since interpretation of each individual image cannot scale to the overwhelming number of rep-

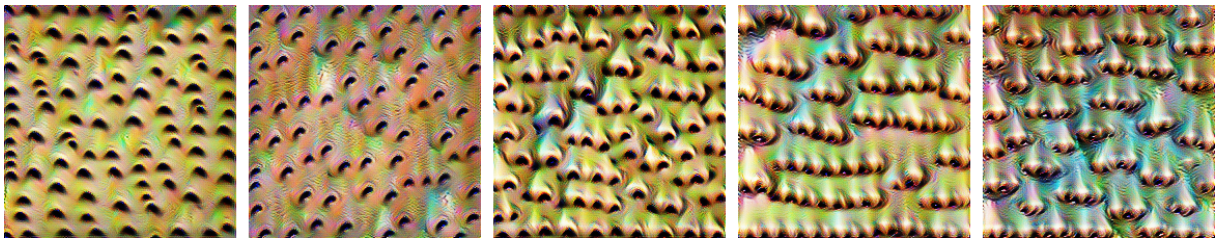

Figure 1.1: Set of synthetic images generated through input optimization. These generated images start as random noise and are optimized to maximize some specific neural pattern in the CNN. Though some of them appear to be wildly different, all samples in this set maximize the same pattern: an isolated channel of an intermediate layer of a face recognition network.

resentational units in modern deep architectures. Some concepts might also be encoded in the network as non-trivial combinations of many filters and neurons, which are unlikely to be perceived when they are visualized individually. Furthermore, such a substantial reliance on human interpretation could hinder the acceptance of any conclusion derived from these methods since it is difficult to precisely estimate the impact of a biased human operator. This is specially concerning because the obtained images are often influenced by strong priors [54; 30] which are usually tuned according to the subjective preferences of the operator.

Figure 1.1 shows a set of synthesized images containing repeated structures resembling nostrils and noses. All images in the figure were obtained through optimization of the same neural pattern in the network but they are undeniably visually distinct from each other. Such a large variance prevents fair comparisons between visualizations of different patterns since it is hard to identify whether some subtle difference between the generated images is caused by an actual distinction in the network representation or just some fluctuation in the generative method. Thus, exclusive reliance on interpretation of visualization methods may also produce misleading conclusions.

The aforementioned limitations illustrate the relevance of a more rigorous and objective method for understanding CNNs. This thesis investigates the extraction of quantitative measurements, instead of images, as a suitable candidate to overcome the subjectivity and comparability issues of visualization techniques. The proposed methodology still does not scale to all possible combinations of neurons and filters, but an objective and replicable procedure is proposed to identify small subsets of these representational units that require further investigation. In a broader context, the relevance of understanding deep neural networks is also reflected in concerns regarding bias and fairness of the optimized models, specially in real-world applications. Face recognition networks are particularly deserving of attention, as they directly affect the life of individuals, with potential to cause disproportionate embarrassment or harm, even if unintentionally, to

particular groups of people [4] if not properly comprehended and regulated.

The face recognition domain also exhibits an important advantage that simplifies quantitative analysis of network features, specially in the more fine-grained examination of individual neurons. In a generic domain, associating semantic meaning to neurons is an arduous task due to their sensitivity to geometric transformations of the input image, specially in the shallower layers of the network. For instance, a simple translation or rotation of the input can dictate whether or not the receptive field of a neuron is able to capture some semantic structure in the image. In this scenario, a neuron may not reliably respond to a stimulus in the image even if it is discriminative of that characteristic, simply because the input is not properly aligned. However, this problem can be mitigated in facial images by lining up keypoints corresponding to the eyes, nose and mouth. Since the facial structure, i.e., relative location of facial parts, is also maintained across most samples in the dataset, we can expect that each neuron will typically observe the same facial region for all aligned input images. This setup cannot be replicated in multi-class datasets due to the absence of a common keypoint to align samples from all different classes. Thus, the scope of this thesis is limited to CNNs optimized under a facial recognition task.

1.2 Objectives

This work addresses the problem of understanding what is encoded across the layers of a face recognition CNN through a quantitative analysis of how several attributes are represented across the network. Specifically, classifiers are trained and evaluated to predict 40 facial attributes while using intermediate representations of three different CNN architectures optimized for face recognition. The extracted representations spam the entire depth of the evaluated networks as well as the granularity levels of neural, channel and entire layer outputs. The achieved accuracies are analysed to reveal the importance of each attribute in the face recognition model and also how the attribute information is distributed among the layers, channels and neurons of the network.

1.3 Contributions

The main contribution of this thesis comes from a thorough analysis of how and which face attributes are encoded in face recognition networks. Our analysis revealed, for

instance, that soft-biometric attributes are better encoded than facial geometry attributes, which may impact the performance of face recognition systems in scenarios where the user may try to conceal his/her identity with apparels such as hats and eyeglasses. We also observed some potential biased representations in the network, which where evidenced by the fact that classifiers for one specific hair color demonstrated far inferior performance than other hair colors. The proposed approach is also able to identify the most relevant neurons and channels in the network for the evaluated attributes. For some attributes, a single neural output from one of the intermediate layers is able to produce a classifier with an accuracy gap bellow 3 percentage points in comparison to a fully supervised approach. While these results suggest that the attribute information is highly concentrated in the network, we also demonstrate that the attribute representation in one layer is more dependant on a large set of neurons of the previous layer rather than a few highly-relevant neurons of the previous layer.

Some of the results obtained in this thesis were published on the proceedings of the *15th IEEE International Conference on Automatic Face and Gesture Recognition* under the title *Face Attributes as Cues for Deep Face Recognition Understanding* [8].

1.4 Organization

The remainder of this thesis is organized as follows. Chapter 2 introduces the most important concepts underlying the proposed analysis. Chapter 3 briefly reviews some of the most relevant and recent works in face recognition and neural network understanding. Chapter 4 describes the proposed approach in detail. Chapter 5 presents and discusses the obtained results. Finally, chapter 6 synthesizes the main contributions and future directions.

Chapter 2

Background Concepts

This chapter briefly reviews the most important concepts underlying the proposed analysis. Section 2.1 summarizes the structure of neural networks, introducing the nomenclature that is utilized to describe different components and representations inside these models. Section 2.2 reviews Partial Least Squares, a discriminative dimensionality reduction technique that simplifies the analysis of the high-dimensional outputs of intermediate layers.

2.1 Neural Networks

Neural networks were initially proposed as a biologically inspired model for understanding information handling systems [35]. The novelty of this approach stemmed from a design in which the information perceived by the brain is never actually compared against any patterns retained in memory. Instead, this model proposes that the responses for the stimuli are automatically generated by the path of existing connections between the neurons in the network. In this scenario, information retention takes place in the form of new or stronger connections inside the network and learning is said to occur when the connections are modified to generate the appropriate response to the stimuli. In the neural network model, this learning process is composed of presenting multiple pairs of input and output patterns to the network, representing the stimuli and the desired responses, respectively. The model will then approximate the desired responses as the connections are updated in the direction that minimizes the error of the output of the network.

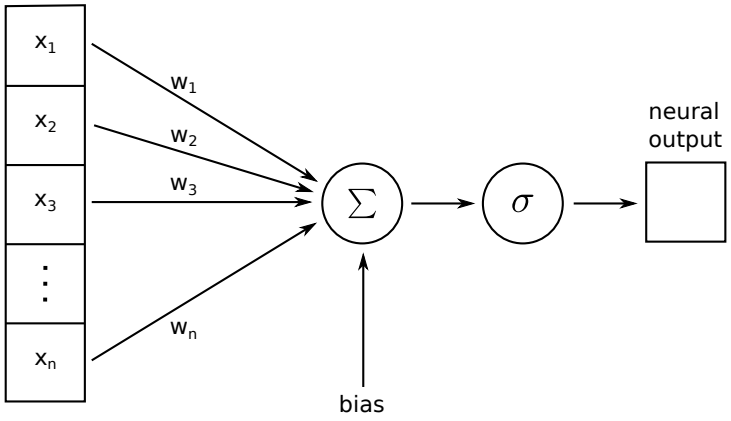

Figure 2.1: Structure of a neuron. The neural output is obtained through the weighted sum of its inputs, x_i, and a bias term, followed by some differentiable activation function. Algebraically, the neural output can be computed by $\sigma(w^T x + b)$, where w and x are column vectors and b is a scalar.

The output of a single neuron in the network is mathematically defined by a simple algebraic sum of its weighted inputs followed by some activation function. The first design of such a neuron, called the perceptron [35], was influenced by the behaviour of the brain, and thus, it was initially restricted to random connections and a threshold activation function to mimic the all-or-nothing response of biological neurons. However, these constraints introduce some important limitations on the model [27], and consequently, they have been adapted in modern implementations. Figure 2.1 shows an up-to-date representation of a neuron. The network was adapted to fully-connect the neuron to all available inputs, including a new parameter called bias, allowing practical computation of the response of multiple neurons through simple algebraic operations. The activation function was also replaced from the threshold function to any differentiable non-linear function, represented by σ. This modification was a crucial step to overcome one of the major limitations of the perceptron: its restriction to linear classification problems.

Theoretically, non-linear problems could be solved with neural networks containing internal representations, i.e., an intermediate set of neurons that provide additional features to the output neurons. The issue was that the perceptron convergence procedure was not powerful enough to update the weights of these more sophisticated networks. A general rule for optimizing networks with internal representations was only proposed decades later [38], relying on the gradient of the an error function with respect to the network parameters. The computation of this gradient is the reason why the activation in modern neural networks moved from the threshold function to any differentiable function.

The gradient rule consolidated the presence of feed-forward layered neural networks in machine learning and pattern recognition research. Figure 2.2 shows a representation of this network, with each neuron being solely represented by its output, omitting both connections and computational logic for improved readability. In this model, the neurons

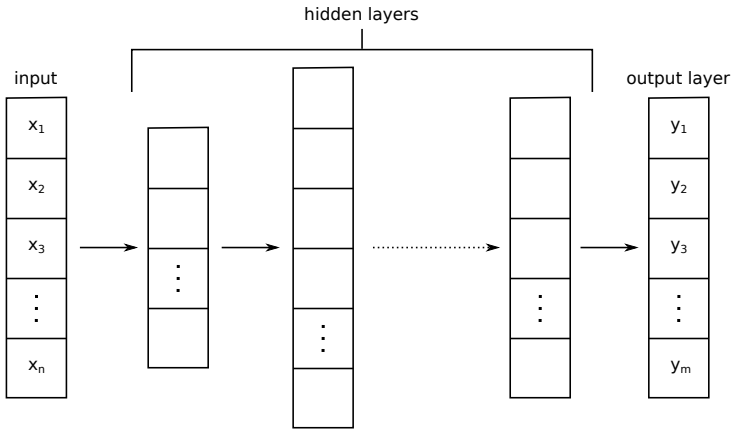

Figure 2.2: Structure of a multilayered neural network. Each neural output is represented by a box, but the neural connections and computational logic are omitted for improved readability. The neural outputs are organized in feature vectors that represent the output of all neurons in the same layer. The arrows indicates that the responses of one layer are used as input in the other layer. The dashed arrow indicates that a sequence with an arbitrary number of layers is omitted.

are organized into a sequence of layers, such that all neurons from the same layer are fully-connected to the responses obtained by the previous layer. Each layer can have any arbitrary number of neurons, which will determine the layer width, or the size of its output vector. The number of layers in the sequence is called the depth of the network, and is usually associated with the network capacity to learn complex representations. Each layer can be categorized as either an output or hidden layer, with the former having direct supervision of an error function, while the latter is only indirectly supervised. The purpose of hidden layers is to uncover discriminative representations of the network input so that the supervised layer can perform complex tasks. The non-linearity property of the activation functions also plays an important role in finding these discriminative representations. This role can be evidenced through a contrast to an alternative network that uses linear activation functions. In such a network, it can be shown that any arbitrarily high number of layers would still provide the same representational power as a single rotation and translation of the input, i.e., subsequent layers would not be able to further distill relevant information from the input. Thus, the effectiveness of neural networks is closely related to their ability to produce complex non-linear representations in their hidden layers.

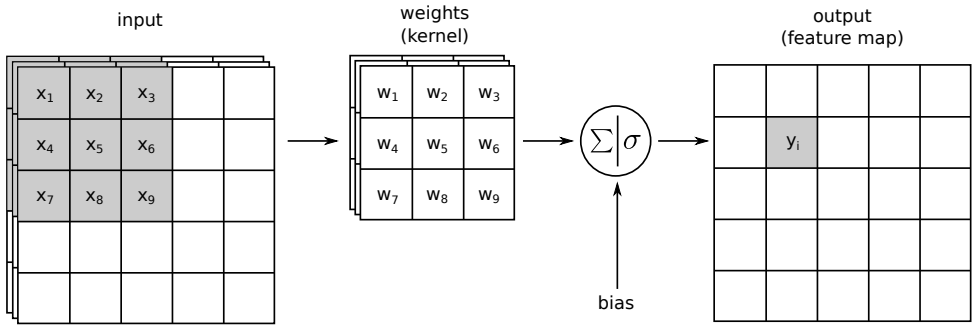

Figure 2.3: Structure of a convolutional layer. The neuron is locally connected to a small region of the input. Both the neural input and output are illustrated by the shaded area. Algebraically, the neural output can be computed by $\sigma(X_i * W + b)$, denoting the convolution between the local region and the kernel matrix plus a bias term.

Even with the discriminative power of hidden layers, the feed-forward layered network design still struggles with high-dimensional inputs such as images, videos and other multimedia data. The fully-connected nature of these networks causes the number of parameters to grow very quickly in respect to the input size. Consequently, when handling high-dimensional data, neural networks were usually employed in combination with handcrafted feature extraction, to first obtain lower-dimensional representations from the original data. However, this strategy is flawed since the extracted features are not precisely optimized to contain all of the relevant information for the target task.

Convolutional neural networks are able to overcome these limitations by exploiting the locality of information in image and video data. Figure 2.3 shows the basic structure of a 2-dimensional convolutional layer. Each neural output is locally connected to a small region of the input, greatly reducing the number of parameters. The convolutional kernel is still able to extract relevant information from the data because most of the correlation between the features occurs between spatially adjacent points. This layer design also exploits the reuse of a single weight matrix in a sliding-window fashion to obtain multiple neural responses over different regions of the input. The neural responses obtained from the same kernel are grouped into a structure called a feature-map. Each convolutional layer may contain multiple kernels, and thus, the entire layer output will be composed over many feature-maps organized across multiple channels.

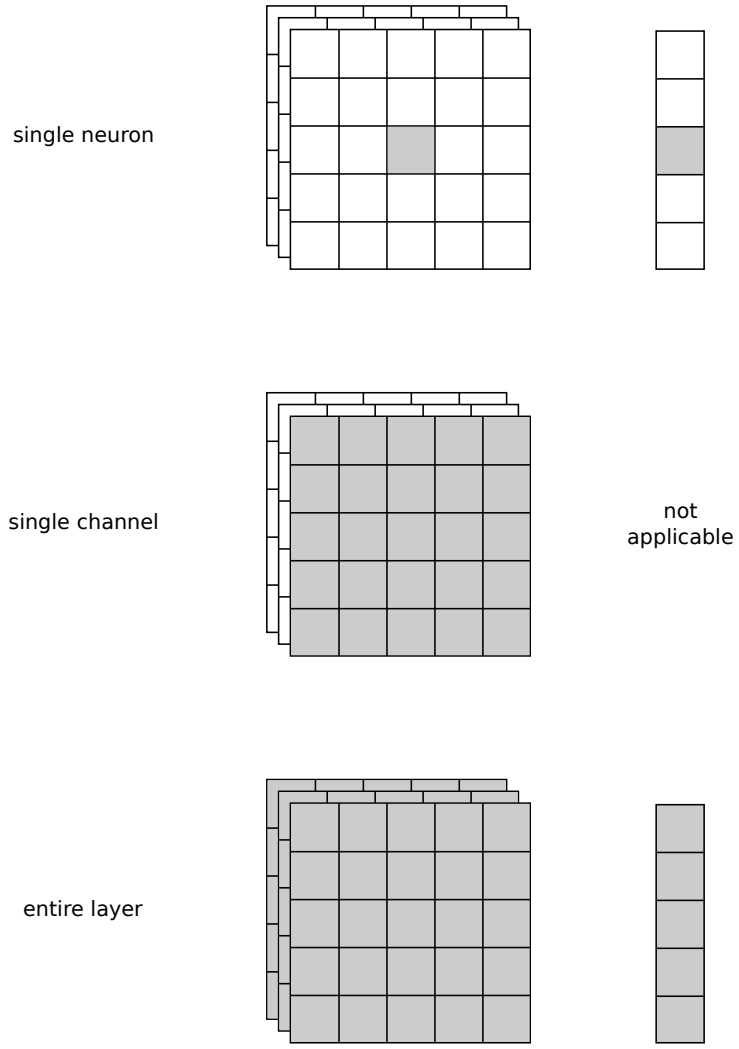

Figure 2.4: Representational units of a neural network. Each representational unit is associated to the correspondent shaded area of the same row. Each row contains two different representations, corresponding to the outputs of 2-dimensional and 1-dimensional layers, on the left and right, respectively.

Generally, neural networks may also contain different types of layers with other operators such as pooling and flattening. However, the output of most of these different layers would still follow the structure described for either the fully-connected or the convolutional layer, depending on whether it outputs a feature-vector or a feature-map. Figure 2.4 shows a summary of the representational units that will be further explored in this work. The neural output corresponds to a single response, i.e., a single floating-point number, in the layer output. The channel output correspond to the response of all neurons that share the same kernel, though this structure does not exist in layers that produce feature-vectors. Finally, the union of all neural responses from the same layer corresponds to the entire layer output.

Modern neural networks may also employ more sophisticated arrangements in their layers rather than the traditional sequential structure. These arrangements can have sub-

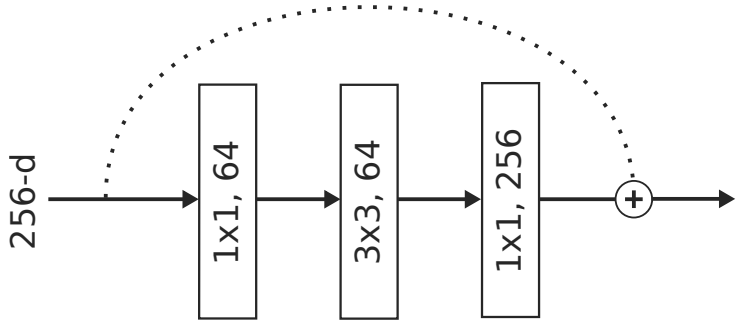

Figure 2.5: Structure of a residual block. The dashed line indicates a shortcut connection.

stantial impact on the network performance, but in essence, they are simply a combination of features from multiple depths and branches in the network. Thus, even the complex modern networks still produce the same feature-maps and feature-vectors that are shown in Figure 2.4. The combination mechanism of these representations may be performed through several operations such as addition, concatenation or multiplication, each with a distinct semantic interpretation.

The residual framework reformulates the network structure such that each layer learns a residual representation as a function of the output of a previous layer. Figure 2.5 illustrates an instance of this formulation which is realized by a shortcut connection, represented by the dashed line, that adds two feature maps of the same dimensionality, The main benefit of this configuration is that it facilitates the training of very deep networks, allowing optimization of networks with thousands of layers. In this work, the shortcut connections will also be examined in detail when analysing the feature maps of all layers.

2.2 Partial Least Squares

The high-dimensionality nature of hidden layers is one of the most challenging aspects of analysing intermediate representations of deep networks. Some layers can create outputs with a dimensionality that exceed a million features. The issues related to these high-dimensional spaces can be avoided with a dimensionality reduction pre-processing step.

Partial Least Squares (PLS) [51; 36; 12] is a dimensionality reduction technique that exhibits several desirable characteristics for the proposed face attribute analysis. The most important of these characteristics is that PLS is able to preserve the discriminative information between predictors and labels, i.e., the layer outputs and face attributes, by maximizing the covariance between these two data blocks. Essentially, PLS decomposes the independent and dependent variables, X and Y into

$$X = TP^T + E, \tag{2.1}$$

$$Y = UQ^T + F, \tag{2.2}$$

where T and U are the scores, P and Q are the loadings, and E and F are the residuals of each block. If optimized separately, these equations are equivalent to the principal components analysis of X and Y. However, PLS is able to retain the discriminative information in the projection through a joint optimization procedure, where a linear relationship between the two blocks is modelled such that

$$U = BT, \tag{2.3}$$

where B is a diagonal matrix of the regression coefficients between the two projection scores.

Algorithm 1 shows the optimization process of PLS through the Non-Linear Iterative Partial Least Squares (NIPALS) [51; 12] algorithm. For visual clarity, we also omit normalization of the resulting vectors in the innermost loop, though it is a necessary step for convergence and numerical stability.

Algorithm 1 NIPALS

Input: $X \in \mathbb{R}^{n \times m}, Y \in \mathbb{R}^{n \times k}, c \in \mathbb{Z}$

$E, F \leftarrow X, Y$
for $i \leftarrow 0$ **to** c **do**
  $u = random\ initialization$
  repeat
    $w \leftarrow E^T u$
    $t \leftarrow Ew$
    $q \leftarrow F^T t$
    $u \leftarrow Fq$
  until convergence of t
  $b \leftarrow u^T t$
  $p \leftarrow E^T t$
  $E \leftarrow E - tp^T$
  $F \leftarrow F - btq^T$
end for

The PLS projection provides several benefits in the context of the analysis the representations of hidden layers. From a computational aspect, the usage of the NIPALS algorithm allows for an efficient projection of the high-dimensional intermediate features since it does not require the computation of the covariance matrix, which would be extremely expensive given the size of the intermediate features. Furthermore, the obtained weights can also be used to determine the relevance of each individual feature X_j in the projection, through the Variable Importance in Projection (VIP) technique [52; 10; 26], which is described by the following equation

$$\mathrm{VIP}(X_j) = \sqrt{m \sum_{i=1}^{k} \frac{SS_i \frac{W_{ij}}{\|W_i\|^2}}{\sum SS_i}}, \tag{2.4}$$

where SS_i denotes the sum of squares explained by the i-th component, which can be alternatively expressed as $q_i^2 t_i^T t_i$. In this work, each feature X_j represents the output of a single neuron, with its VIP score determining its relevance to the label representation.

In summary, the high-dimensional layer outputs can be analysed in a more straightforward fashion by first projecting them into a lower-dimensional space with PLS. Then, the analysis may also be extended to include the study of a few individual neurons and channels of each layer, by first measuring which of these units are more relevant in the PLS projected space.

Chapter 3

Related work

This chapter presents some of the relevant works related to this thesis. Section 3.1 briefly reviews the usage of CNNs for face recognition and attribute classification. Since this work is only tangent to these tasks, the related methods are not profoundly detailed in this section. Section 3.2 reviews the modern techniques for CNN understanding, including both visualization and quantitative techniques.

3.1 Face Recognition and Attribute Classification

Deeply-learned face representations [46; 6; 5] coupled with large scale datasets [19; 3] significantly improved the performance of face recognition methods. These datasets contain thousands of identities to fuel the optimization process of increasingly deeper [21; 14] convolutional neural networks. Intermediate layers of these networks are said to contain deeply-learned representations, which are robust and discriminative descriptors, especially at deeper layers. Most of the recent effort on face recognition research has been dedicated to improving these descriptors, mainly by adjusting several loss functions [41; 50; 22; 57; 6; 49; 5; 53; 16] to increase inter-class and reduce intra-class distance in the final embeddings of the network.

Interestingly, the discriminative power of face recognition descriptors can also be improved through joint optimization of identities and facial attributes [34], suggesting that the encoding of these attributes are relevant for the face recognition task. Thus, a reasonable hypothesis is that, even with no explicit joint supervision, face recognition CNNs may learn to represent these attributes implicitly since they encode relevant information for the recognition objective. Our approach uses this hypothesis to probe face recognition networks and reveal how several face attributes are encoded throughout their layers to improve our current understanding of the recognition models. Generally, the techniques applied to attribute recognition are mostly concerned with the interactions between the attributes and how to properly sample under-represented attributes [23; 37; 13].

While our proposed approach is closely related to both face recognition and attribute classification, it is not concerned with improving the state of the art on these tasks. Instead, the main goal is to understand how the face attributes are encoded in the face recognition network.

3.2 Understanding Convolutional Neural Networks

Drawing inspiration from neuroscience, several methods have been proposed to understand CNNS by visualizing the preferred stimuli for specific sets of neural activation [31]. These stimuli are represented by synthesized images obtained from either inversion or maximization of some particular pattern in the network through a variety of techniques such as gradient optimization, deconvolutional [55; 59], up-convolutional [9] and generator networks [29; 28]. Naive generation of these images leads to unintelligible results due to excessive high-frequency patterns and unnatural colors [31; 32]. These issues are usually diminished by modifying this process to include priors for local [54; 25] and global structures [30], resulting in the generation of more interpretable images.

Several limitations of these techniques can be traced back to this objective of generating interpretable visualizations. For instance, the established priors are hand-designed functions whose hyper-parameters are tuned according to the subjective notion of what the user considers to be the most interpretable. The ideal hyper-parameters may also change for different layers, architectures or datasets [31], preventing fair comparisons of different results. These comparisons are also hard to scale, since they require human interpretation of each generated image, which is not viable if an image is to be synthesized for each neuron, filter and layer in the network. Requirement for human interpretation also poses another disadvantage for these methods, as users are able to find meaning in visualizations even when they are generated by random directions of the original neural patterns [48]. In summary, there is a high degree of uncertainty in the results of visualization methods, mainly due to the large influence of the biased interpretation of human operators, and strong regularization priors, that may produce a misleading, but visually pleasing, representation of the network stimuli.

Some of these limitations can be addressed through the extraction of quantitative measurements from the generated images. For instance, reconstruction error and classification consistency can be extracted from images generated by the inversion of a neural pattern [25], quantitatively assessing the robustness of different techniques and regularization priors. Another example can be seen in attribution heatmap techniques, a class of methods that are able to identify the salient region in a real world image with respect to

some neural output [60; 61; 43]. Qualitatively, the visualizations of the salient regions can be inspected by users to determine whether the network is employing the correct semantic structure in the input image to make its predictions. However, these regions can also be systematically evaluated in localization, detection or segmentation protocols, providing a rigorous but straightforward framework to understand different network representations. These quantitative experiments revealed, for instance, that more object detectors emerge in the original representations of the network than in their random directions [2], which emphasizes the importance of the quantitative analysis, since, if not for the quantitative aspect, one could potentially mistakenly conclude that these representations are equivalent once object detectors emerge from both of them.

Extracting quantitative information from the visualizations is an important step towards a more systematic understanding of deep CNNs. However, the quantitative approaches do not necessarily need to be performed on the visualizations, and can, instead, be applied directly on the intermediary representations of the network. Some basic approaches are able to reveal important properties of the network through the extraction of simple statistics from the neural activation patterns. For instance, examination of the last layer of a face recognition CNN revealed that its L2-norm encodes how hard it is to recognize the input image [33]. Furthermore, it has been demonstrated that some neurons of the last layer are highly discriminative of some face attributes or even specific identities [47; 11; 23]. However, the simplistic nature of these experiments limits their applicability to low-dimensional layers, which is usually only represented by the last few layers of the architecture. Analysis of the intermediary layers, however, are usually associated with more sophisticated approaches, frequently focused on understanding the representation of a pre-defined set of attributes across these layers. One such approach is to measure the mutual information between a set of attributes and different hidden layers of the network [7]. Similarly, a set classifiers can be trained to predict these attributes so that their performance or weights [1; 20] can be used to determine how the concept is encoded in different parts of the deep model. While our approach also relies on the performance of classifiers to understand deep networks, it does so with a more extensive set of attributes, in the less explored domain of face recognition, and across different units of representation such as neurons and filters.

Chapter 4

Proposed Approach

This chapter presents the proposed technique to analyse the representation of several face attributes across the layers of a deep face recognition network. Section 4.1 describes a method to probe intermediate layers of the network. Section 4.2 describes how an extension to this method to also analyse fine-grained representations of the layer, such as individual neurons and channels. Figure 4.1 shows a brief summary of the proposed method.

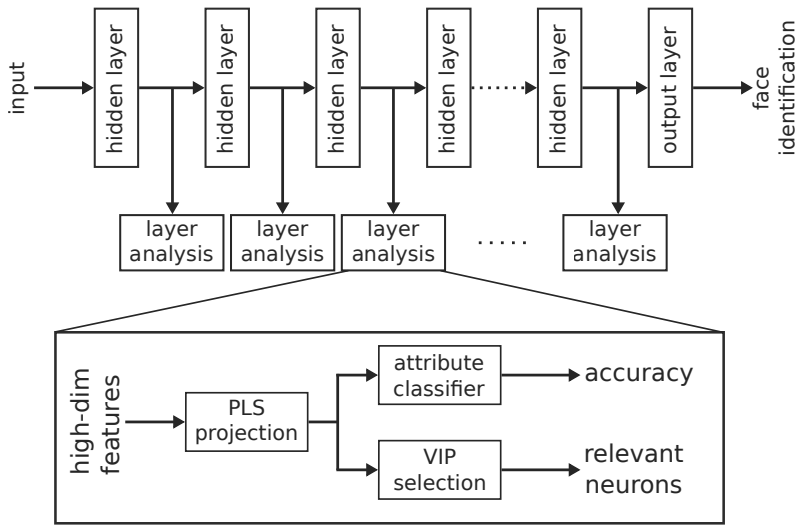

Figure 4.1: For each attribute, we learn a discriminative low-dimensional projection of the output of the hidden layers of a deep network. This low-dimensional projection is used to train a classifier whose accuracy indicates the discriminability of the layer with respect to the evaluated attribute. The optimized weights of the PLS projection is also employed to analyse how the attribute information is spread inside the layer through the scores of a variable selection technique.

4.1 Layerwise Attribute Analysis

The proposed approach examines how each face attribute is encoded across the layers of a face recognition CNN. To this end, an attribute classification pipeline is attached to the output of each layer in the network. This pipeline is trained with a set of labelled images in a supervised setup to predict the occurrence of several facial attributes using the output features of the analysed layer. The accuracy achieved with each classifier indicates to what extent each attribute is represented at different depths of the model, characterizing the layers of a face recognition CNN in a quantitative manner according to its predictive power of the facial attributes, reducing the need for subjective interpretation of the results. Thus, distinctions between layers, architectures or attributes can be measured in a precise, replicable and comparable procedure.

The main assumption of the proposed approach is that the layers containing suitable representations for an attribute will produce a more accurate classifier for that attribute. However, robust classification models and pipelines, such as deep networks, are able to distill richer representations that were not directly encoded in the examined layer output, obtaining a potentially misleading accuracy. Thus, it can be argued that the performance of classical learning methods provide a more faithful portrayal of each layer, since they operate more directly on the analysed features. Attachment of these classifiers is straight-forward for one-dimensional outputs, such as fully-connected and global pooling layers, but convolutional layers require a pre-processing stage to obtain a flattened vector of size $h \times w \times c$, denoting the product of the height, width and number of channels of the convolutional layer.

These flattened convolutional layers can generate extremely high-dimensional outputs. For instance, the convolutional first layer of the VGG16 [45] architecture, generates a $224 \times 224 \times 64$ feature map, which exceeds three million features. Some learning methods are not even applicable in this scenario as the number of features greatly exceeds the number of samples available for training the classifier. While it is theoretically possible to train some other classical methods, they would still be constrained by the curse of dimensionality [18], a phenomenon that causes deterioration of the classification model due to over-fitting on noisy features. Besides the poor performance, there are also practical issues associated with high-dimensional spaces, such as unreasonable computational cost and excessive memory consumption.

Dimensionality reduction techniques are suitable candidates to overcome these issues. The low-dimensional output of these techniques enables the usage of essentially any simple classifier with a much smaller set of labelled samples for each attribute. Since the goal of the analysis is to understand the representation of the facial attribute, the obtained low-dimensional transformation must be able to retain the information about the evaluated attribute. This can be achieved with Partial Least Squares, a technique that is able to maximize the covariance between the network features and attribute labels in the projected space [51]. Furthermore, since the projection obtained with PLS is essentially a subspace from a rotation of the original features, and thus, it can be argued that the levels of semantic information is maintained across this transformation and that the classifier performance would still be a realistic indicator of the attribute representation in the original feature.

In its essence, the proposed approach projects, into the low-dimensional space, the layer output $X \in \mathbb{R}^{n \times m}$ and the labels for one facial attribute $Y_i \in \mathbb{R}^{n \times 1}$, where n denotes the number of samples, and m, the number of features in the original space. Each layer output is projected onto multiple spaces such that each facial attribute is associated with one low dimensional projection, T_i, refraining from projecting all attributes into a single subspace. From a theoretical perspective, this decision avoids interference between the attributes in the projected space, and also allows a fine-grained analysis for each individual attribute, which will be described in detail in Section 4.2. From a practical perspective, mutual projection of all attributes would also be prohibitively expensive, since it would require a larger number of samples n to obtain enough variance to describe all attributes. For instance, a good selection of samples for the *Bald* attribute is not likely be have enough variance to describe the *Gray Hair* attribute. Selecting a large and diverse enough set of samples for all attributes leads to computational issues due to the large size of the input matrix X with the features in the original space. However, this issue is avoided when a smaller and unique set of samples is selected to optimize separate projections for each attribute.

The final classification pipeline consists of projecting the layer output X onto a low dimensional projection T_i, which is then used to train a classifier to predict Y_i. Quadratic Discriminant Analysis (QDA) is employed as the classifier, following the setup of [42], though any other classifier could likely be employed without greatly impacting the obtained results. Since QDA is a non-linear classifier, it is able to extract some extra information from the linear projection obtained with PLS, which leads to slightly better accuracies. The method could also be easily extended to analyse continuous concepts, such as facial pose, by simply replacing the classifier with a regression method.

In summary, the proposed analysis consists of executing the exact same classification pipeline while modifying either its input features X, or its target Y_i. Figure 4.2 shows a visual representation of the approach. The approach may be easily interpreted in

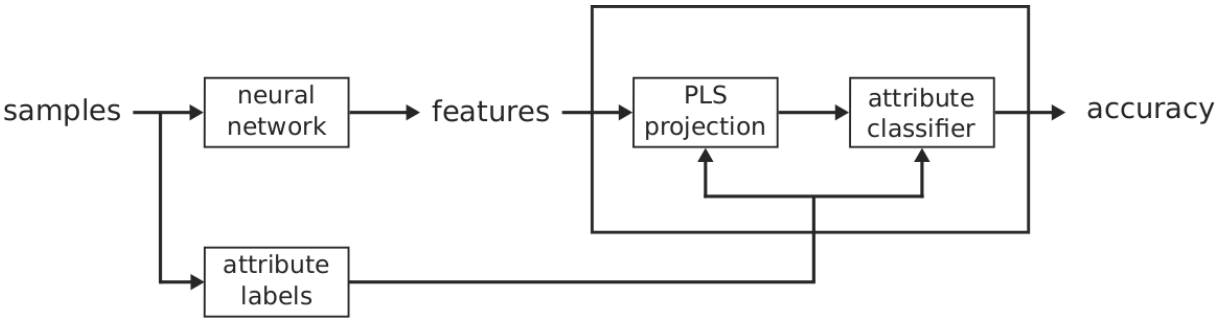

Figure 4.2: Layerwise proposed analysis. The proposed approach extracts accuracy measurements from the neural network layers. Different layers can be analysed by changing which features are being extracted from the network. Different attributes can be analysed by changing which samples and labels are available for the optimization of the classification pipeline.

an analogy to a multimeter measurement tool. While the multimeter is able to measure electrical properties such as voltage, resistance and current, the proposed approach will measure how different facial attributes are encoded in the network by simply changing which attributes of the samples are being labelled. Similarly to how the multimeter can be connected to different components of a electrical circuit, our also approach may also be connected to any of the layers or feature of the deep network, revealing the attribute is encoded at different network depths. Naturally, different networks architectures can be analysed by replacing the neural network components with the desired network model. Thus, this rigid and simple pipeline allows fair comparison of each obtained measurement, since we can precisely control which representations and attributes are being evaluated with the pipeline.

4.2 Fine-grained Attribute Analysis

In addition to understanding the representation of the facial attributes over the layers of the network, the proposed approach is also able to obtain a more fine-grained analysis of individual neurons and filters in the network. This extension is as straightforward as changing the input features of the classification pipeline to receive the outputs of these representational units instead of the entire layer. The key issue is that training and evaluating a classifier for each such unit would be a prohibitively expensive procedure for any modern deep neural network architecture.

This issue can be avoided by evaluating the pipeline in a small subset of these units. This setup relies on the assumption that there exists an association between the facial attributes and individual neurons and filters in the network. Following this assumption, only the units that are relevant for the representation of an attribute would need to be evaluated in the classification pipeline. The relevance of each neural output for a specific attribute can be determined with the Variable Importance in Projection (VIP) technique [52; 10; 26]. This technique scores each individual feature X_j, corresponding to a neural output, according to its influence in the PLS discriminative projection. The top-scoring neurons can then be evaluated through the pipeline by simply replacing the entire layer output with the desired selection of neural outputs. Similarly, the top-scoring filters can also be identified by averaging the importance of all neurons which share the same weight, i.e., the neural outputs that belong to the same channel. Thus, the proposed approach is also able to determine how the attribute information is distributed inside each layer of the network by comparing the discriminative power of small selections of these representational units.

Figure 4.3 summarizes the entire approach, with the top part representing the additional steps used in the fine-grained analysis proposed in this section. Essentially, the PLS projection obtained with the entire layer, which was described in the previous section, can be used to determine the VIP score of each feature in the original space. Since each individual feature represents the output of a neuron, we can select a few individual neurons by selecting the features with higher VIP score. Similarly, the VIP score can also be aggregated to select entire channels instead of neurons. This selection can then be processed by a similar classification pipeline, though in the case of the selection of neurons, there is no need to project them into a smaller subspace, since they are already low-dimensional, as indicated by the dashed lines in the diagram around the PLS projection.

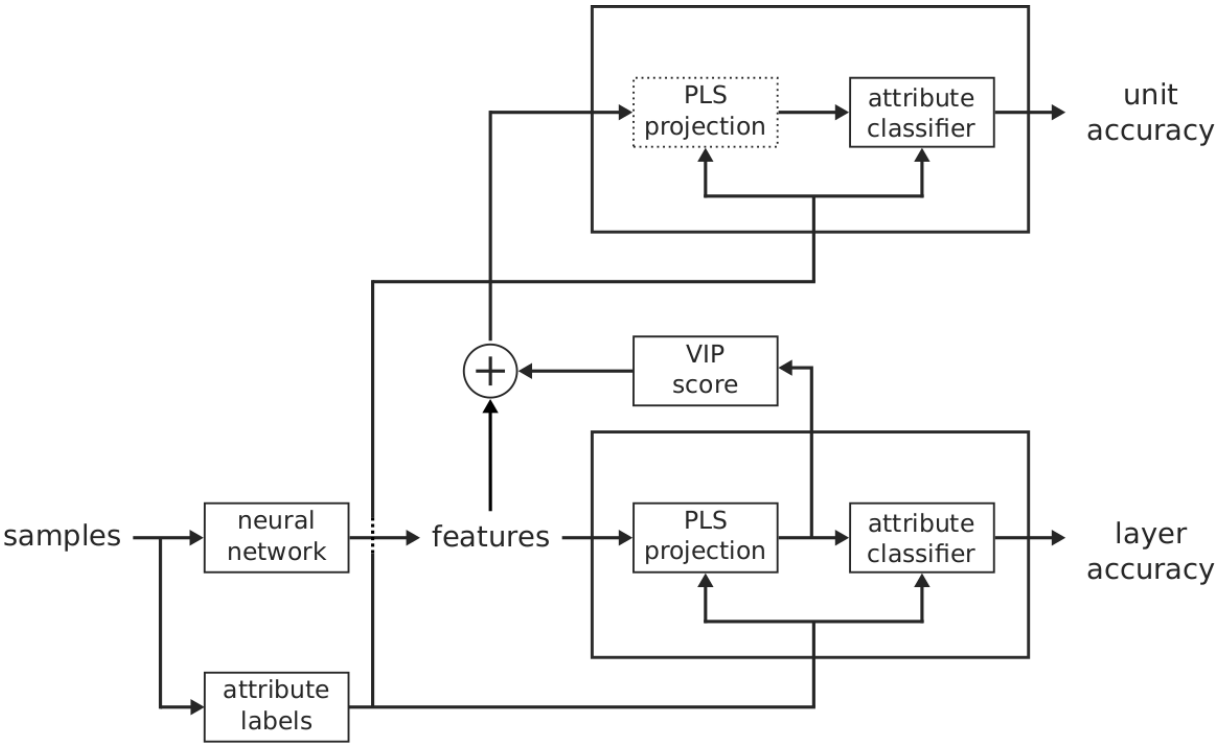

Figure 4.3: Neuronwise and channelwise proposed analysis. The proposed approach is able to analyse any feature or combination of features of the network, including individual neurons and channels. Since there are too many of such individual units, we propose to analyse only a subset of these units which are selected according to their VIP score.

In summary, the proposed approach allows the analysis of any attribute across the representations of a deep neural network. Each representation is associated with an accuracy score which allows direct comparison of attributes across multiple layers of the network. These comparisons can also be made between the neural, channel and layer representations with the aid of the VIP score to select a small subset of these representations for comparison.

Chapter 5

Experimental Results

This chapter presents and discusses the results achieved with the method proposed in Chapter 4. Section 5.1 summarizes the setup necessary to reproduce the proposed experiments. Section 5.2 analyses the representation of several facial attributes across the layers, channels and neurons of face recognition deep networks. Section 5.3 shows a more in-depth analysis of the VIP-score, revealing how the attribute information is distributed inside the layers.

5.1 Experimental Setup

This section describes the experimental setup regarding the optimization of face recognition networks and the attribute classifiers. The face recognition networks are the object of study in which our proposed approach analysis is performed. Thus, we detail the settings in which they were optimized both so the analysis can be replicated, as well as to clearly delineate the settings in which our discussion and conclusions are applicable. Our goal is also not to obtain perfectly optimized networks with hyperparameter searches or other convoluted training schemes. Instead, our aim is to guarantee that the analysed face recognition CNNs are optimized under similar settings, so that our results are not distorted by different data-augmentations, datasets, and other possible sources of variance that would come from analysing pretrained models available in the literature.

5.1.1 Face Recognition Networks

The analysed face recognition CNNs are trained from scratch to remove possible sources of variance such as different training datasets, loss functions and augmentation procedures. The networks are trained and validated on VGGFace2 [3] and LFW [15] datasets, respectively, and the performance on the validation set is used to select the best weights.

The training protocol roughly follows the implementations details described by [3]. The identities are uniformly sampled during the optimization process, and each sample has a 20% chance of being to transformed to grayscale. Each image is also resized such that its shortest side has 256 pixels, and then randomly cropped to a square region with length of 224 pixels. Stochastic gradient descent is used with an initial learning rate of 0.005, which is decreased tenfold whenever the error function plateaus. The networks are optimized under the softmax criterion with a batch size of 64, distributed among 2 GPUs, for 50 epochs.

Networks of three different architectures were optimized following this protocol, namely Resnet50 [14], MobilenetV2 [40] and VGG16 [45]. Each architecture was optimized at least three times from different random initializations, and for the remainder of this work, reported results of each architecture will represent a mean performance, unless stated otherwise. Table 5.1 shows the obtained accuracy on identification protocol of VGGFace2, demonstrating that the evaluated CNNs achieved a performance that is comparable with the results reported in the literature. While these results may not be optimal, they are adequate enough to demonstrate that our analysis is being performed on face recognitions CNNs with a performance that is up to date with the state of the art.

Table 5.1: Accuracy on the identification protocol of VGGFace2 test split

Architecture	Accuracy (std)
Resnet50	92.44 (0.16)
MobilenetV2	90.54 (0.33)
VGG16	90.13 (0.53)

5.1.2 Face Attributes

A total of 40 binary face attributes from the CelebA [23] dataset were employed to probe the intermediate layers of the optimized face recognition CNNs. No additional pre-processing is employed since the original images of the dataset are already aligned through facial keypoints. The entire training split cannot be loaded into memory for PLS optimization due to the high-dimensional nature of the intermediate features. Thus, the original training and validation splits are combined into a single pool, from which 2048 and 6144 samples are drawn for training and validation, respectively. Each classifier received a unique balanced set of samples, otherwise it would be difficult to establish a single set of 2048 images with enough variance among the attributes to train all classifiers. The test split was not modified so that our final results can be compared with other methods. Unless specified, all reported results were obtained on the customized validation split.

A classification pipeline consisting of PLS projection and Quadratic Discriminant Analysis (QDA) is trained for each attribute and layer of interest. The VIP score is computed for each optimized PLS model to identify 8 top-scoring neurons and channels for each attribute. New instances of the PLS+QDA pipeline are trained with the top-channels representation, with one of the instances using all top-channels as input while other instances use each individual channel. An equivalent setup is used for the top-neurons representation, though PLS projection is not employed in this setup since the top-neurons already have the same dimensions as the target projection of 8 components. While there is a strong motivation for using PLS in the pipeline, QDA could be replaced by any other simple classifier without drastically changing the obtained results.

5.2 Attribute Analysis

Figures 5.1 to 5.3 show the accuracies achieved by each classifier over the depth of the CNNs optimized for face recognition. The horizontal axes of the plots are adjusted according to the maximum depth of the network that is associated with each figure. Each individual result is composed of three different curves with colors blue, green and orange, which denote the accuracies obtained by the classifier with the entire layer representation, the selection of channels and the selection of neurons, respectively. The green curve is slightly shorter than its counterparts because the outputs of the last layers are one-dimensional, and thus, cannot be examined with a channelwise approach. Throughout this section, subsets of these charts will be presented in other figures to highlight specific patterns that are harder to visualize amidst the excessive amount of information available in this initial overview. However, these figures still serve the purpose of providing a reference for all obtained results as well as allowing the observation of general trends across all attributes.

One observation that can be made from Figures 5.1 to 5.3 is that most of the evaluated classifiers are able to exceed 90% accuracy, which corroborates the hypothesis that many face attributes are well represented across the layers of face recognition networks. The representation of each attribute also seems to be condensed in a small selection of channels, as indicated by a small accuracy gap between the green and blue curves throughout all of the depth of the network. This phenomenon is also observed for the selection of neurons from deeper layers of the network, reinforcing the notion that the facial attributes information are concentrated across few representational units of each layer. However, the top-neurons representation from shallower layers exhibit a much wider accuracy gap for most of the evaluated attributes, which may be explained by their small receptive fields. Even with aligned facial images, small receptive fields may still not be able to consistently capture the relevant region of input due to peculiarities of each individual, such as hairstyles and facial expressions, that could cause small changes in the position of facial parts. This hypothesis is also supported by fact that highly accurate classifiers can be produced with a selection of channels from these same shallow layers. The only distinction between the top-channels and top-neurons representations is that the former convolves a filter over all valid spatial locations of the input, while the latter is restricted to a single patch. This observation indicates that the convolutional kernels in the shallow layers are capable of producing discriminative outputs provided that they are able to process larger regions of the input, which could be achieved either with wider receptive fields, as in the deeper layers, or with the processing of multiple regions of the input, as in the channels representation.

Another interesting observation is that, for the entire layer representation, most of

Figure 5.1: Attribute prediction accuracy over the depth of the Resnet50 network. The blue curve measures the accuracy obtained by an attribute classifier that uses the entire layer response as its input. The green and orange curves replace the layer representation by a subset containing the eight top-scoring filters and neurons, respectively.

Figure 5.2: Attribute prediction accuracy over the depth of the MobilenetV2 network. The blue curve measures the accuracy obtained by an attribute classifier that uses the entire layer response as its input. The green and orange curves replace the layer representation by a subset containing the eight top-scoring filters and neurons, respectively.

Figure 5.3: Attribute prediction accuracy over the depth of the VGG16 network. The blue curve measures the accuracy obtained by an attribute classifier that uses the entire layer response as its input. The green and orange curves replace the layer representation by a subset containing the eight top-scoring filters and neurons, respectively.

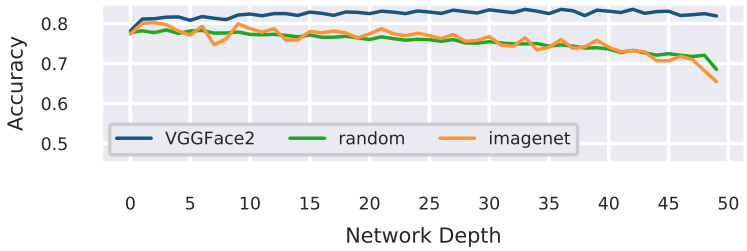

Figure 5.4: Mean attribute accuracy with different optimization setups using the entire layer representation. Each curve denotes the mean accuracy of all attributes of a Resnet50 network optimized with different datasets, or not optimized at all (random).

the curves exhibit either a plateau or very subtle improvements after the first few layers of the network. Since these initial representations are not closely related with the objective function, this phenomenon may challenge the assumption that the obtained attribute representations are produced by the face recognition optimization procedure.

Figure 5.4 compares the mean attribute accuracy of a Resnet50 network under three different optimization setups: random initialization, face recognition, and image classification [39]. All three setups have a roughly equivalent accuracy at the initial layers of the network, which indicates that a high accuracy at the shallower layers is more related to the simplistic nature of the evaluated attributes rather than the network optimization procedure. However, at each subsequent layer, the average codification of the attributes slightly improves in the face recognition network, but worsens in the other two networks, and thus, it seems that the accuracy obtained with deeper representations is indeed determined by the optimization procedure.

Most of the evaluated attributes exhibit an initial accuracy of roughly 80%, with the exceptions generally belonging to categories of higher complexity, encoding age, hair texture and facial geometry. In the obtained results, these categories are represented by binary attributes such as *Young*, *Wavy Hair*, *Straight Hair*, *Big Nose*, *Pointy Nose*, *Big Lips* and *Oval Face*. Surprisingly, the gender attribute, *Male*, has one of the highest initial accuracies, even though it is associated with a higher-level concept that is expected to be well encoded only at the deeper layers. However, it may be the case that the gender attribute classifier is exploiting some low-level attribute that is highly correlated with gender, such as the thickness of the eyebrows or presence of facial hair.

The *Brown Hair* attribute also exhibits a low accuracy in the shallower layers, though this is an anomalous behaviour among the attributes that encode hair color. Figure 5.5 shows the accuracies obtained by the classifiers with the entire layer representation for five semantically similar attributes, with four of them encoding a hair color. The accuracy gap that is present in the initial layers is essentially maintained throughout all

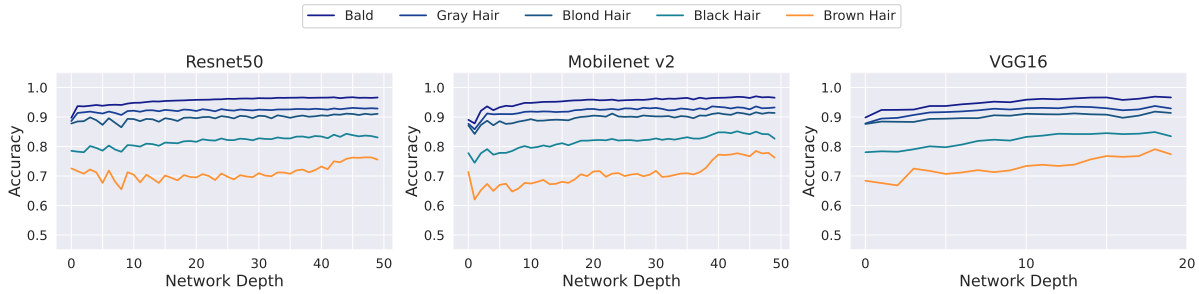

Figure 5.5: Accuracy of facial attributes encoding baldness and hair colors using the entire layer representation. Even though the brown hair color is semantically equivalent to the other colors, it exhibits a widely inferior accuracy. Best viewed in color.

of the depth of the network, often ranging between 10 to 20 percentage points below its counterparts. Besides this large gap, the anomalous behaviour of brown color is also manifested in its unsteady accuracy curve over the network depth. While the other attributes demonstrate a considerably steady accuracy improvement, the brown hair curve exhibits an unstable behaviour, with large oscillations in accuracy, as well as with the most significant accuracy improvement concentrated in the last few layers. The obtained results do not provide enough information to establish what causes this distinguished behaviour with the brown hair color. This explanation could be a combination of multiple factors, such as a bias in the face recognition dataset used to train the networks, or perhaps, a larger variance of hairstyles in brown hair individuals in the face attribute dataset.

Nonetheless, this phenomenon suggests that specific variations of a trait, such as the hair color, are encoded differently in the network. In turn, this could cause the face recognition system to favor particular characteristics and thus, potentially neglect particular ethnic groups. Even though the proposed approach is not able to provide an explanation to this possibly biased behaviours, it can at least identify these scenarios more reliably than visualization techniques. Then, researchers and developers may actively investigate each specific deficiency in the representation that was identified by the proposed method.

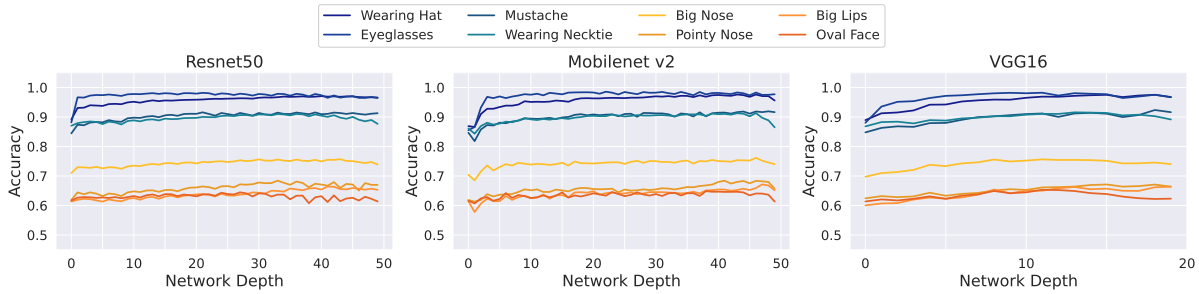

Figure 5.6: Accuracy of facial attributes encoding primary (blue tones) and soft (orange tones) biometrics using the entire layer representation. Best viewed in color.

Another interesting perspective on the obtained results comes from the contrast between primary and soft biometric attributes. Soft biometrics are characteristics of the individual that lacks either distinctiveness or permanence [17], which makes them less reliable for identification since they are not likely to remain constant between all image acquisitions or even be unique to a single individual. Primary biometrics are intrinsic attributes which are particular to each person such as the fingerprint, iris, face geometry or shape of a body part. These attributes are less likely to change between different photo acquisitions since they can only be altered through invasive procedures such as plastic surgeries. Considering that the face recognition optimization must classify hundreds of samples of the same individual, it is expected that the network would rely on characteristics that are more likely to be shared among all samples, i.e., the primary biometric attributes.

Figure 5.6 shows the accuracies obtained for the all attributes encoding primary biometrics, in orange tones, and a selection of soft biometric attributes, in blue tones. While soft biometrics are able to consistently reach 90% accuracy, even the best represent primary attribute barely reaches 75%. Essentially, the attributes enconding primary biometrics showcase the worst performances among all evaluated attributes. This counterintuitive result suggests that the performance of the evaluated face recognition network is more dependent on visual characteristics that could be easily modified, such as wearing different acessories or changing the hair style. This phenomenon could negatively impact the performance of face recognition systems in scenarios in which there is no guaratee of user cooperation, or in which the probe and gallery image sets require different dress codes, e.g., photos from official documents in contrast to in-the-wild image acquisitions.

Some of the evaluated attributes encode information about the photo acquisition process or facial pose. In this discussion, these attributes will be referred as non-biometrical characteristics since their manifestation is not correlated with specific individuals. Figure 5.7 show the accuracy curve of these attributes. It should be noted that

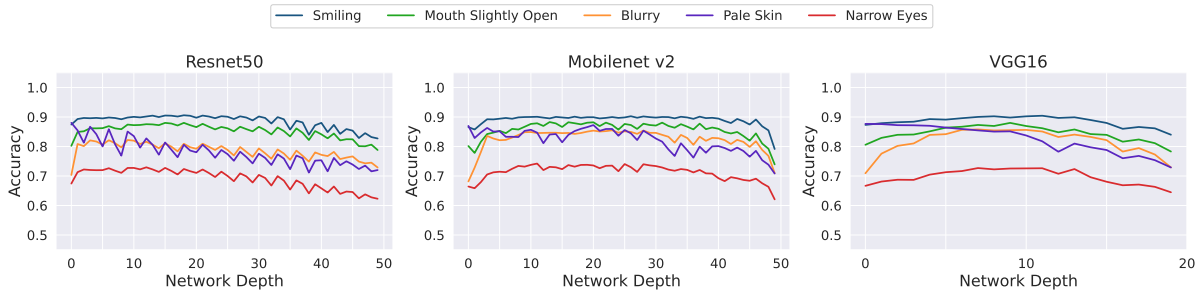

Figure 5.7: Accuracy of facial attributes encoding non-biometric attributes using the entire layer representation. Best viewed in color.

the attribute *Pale Skin* is more associated with make-up style or increased brightness conditions rather than light-complexion skin tone, and that the attribute *Narrow Eyes* is encoding the facial expression of squeezing the eyelids together, rather than the general shape of the eyes.

The accuracy of non-biometrical attributes generally decreases over the depth of the network. This suggests that the face recognition embeddings discards the information of these attributes, since they are not relevant for identification of the images. One unusual aspect of these accuracy drops is that they follow a distinctive sawtooth pattern both in the MobilenetV2 and Resnet50 architectures. While this shape may initially seem to be random noise, it does, in fact follow a meticulous arrangement closely associated with the structure of the residual blocks [14] of these architectures. Fig. 5.8 illustrates the relationship between the accuracy curve of the *Narrow Eyes* attribute and the residual block of the Resnet50 architecture. Essentially, for each residual block, we can systematically identify a local maximum after the skip connection, and a local minimum at the intermediate residual representation. In fact, these local maxima and minima are present in all of the evaluated attributes, though more subtly in the attributes that are relevant for the recognition task, and more evidently in the non-biometrical traits, which exhibit sharp accuracy decreases at their residual representations. When the residuals are combined through the skip-connection, they negatively impact the performance, indicating that the residual information is subtracting the non relevant information from the deeper representations.

One limitation still remaining in the analysis is the comparison of attributes that are not semantically similar. Two attributes that achieve the same accuracy do not necessarily have a representation of equivalent importance in the network. For instance, it could be the case that the evaluation set of one of the attributes contains more noisy labels which would negatively impact the accuracy of the classifiers. Thus, using the absolute classifier performance may be an unfair metric to determine which attributes

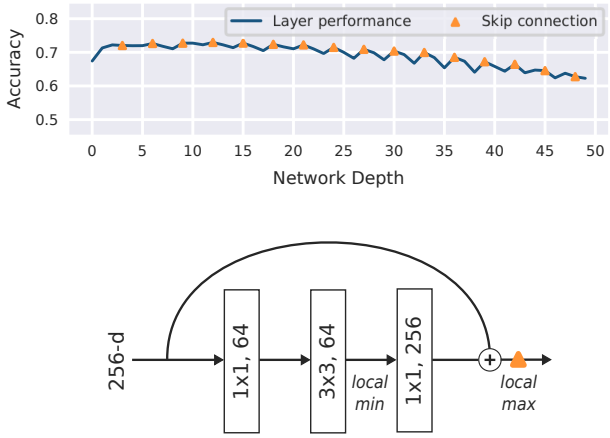

Figure 5.8: Top: the entire layer performance of the Resnet50 network for the *Narrow Eyes* attribute. Bottom: residual block of the Resnet50 architecture. The sawtooth pattern is caused by the skip connections and bottlenecks of the residual block, with local maxima coinciding with the representation obtained after the skip connection.

are well represented in the network. However, this issue can be avoided by comparing the accuracies obtained the proposed analysis against fully-supervised approaches for face attribute prediction. In this manner, it is possible to determine if an attribute is well represented in the face recognition network if the obtained accuracy is close to that of a representation that was supervised for this task.

Table 5.2 compares the accuracy obtained in the analysis with a multitask Resnet50 baseline that was optimized to jointly predict all 40 attributes. The table is ordered according to the relative accuracy between the evaluated classifiers and the baseline, with the best-performing attributes appearing on the top. The reported accuracies were obtained with the original test set of CelebA [23], and correspond to the performance of the best classifiers for each given representation in the validation set. Contrary to the previous analysis, the reported results for channel and neuron representations refer to a single channel or a single neuron instead of the top-8 units that was previously used. In general, many attributes were able to produce results that are almost as accurate as the multi-task approaches, demonstrating that the face recognition network encodes them very accurately.

In Table 5.2, attributes such as *High Cheekbones* and *Wavy Hair*, are not able to achieve highly accurate results, but they are still comparable to the accuracy obtained by the supervised multitask network. Thus, while the absolute accuracy values may suggest that some attributes are not well encoded, the relative accuracies reveal that the face recognition representations are almost as discriminative as if they were optimized for to predict these attributes. Generally, the top performing attributes correspond to soft biometrics while the worst performances correspond to the primary biometrics, reaffirming the hypothesis that face recognition CNNs favours the representation of the former. The non-biometrical attributes *Smiling* and *Mouth Slightly Open* are also able to achieve highly

Table 5.2: Accuracy for the best (L)ayer, (C)hannel and (N)euron

	Resnet50			Mobilenet V2			VGG16			Multitask
	(L)	(C)	(N)	(L)	(C)	(N)	(L)	(C)	(N)	
Male	97.7	96.4	96.2	98.2	97.4	97.3	97.9	90.8	89.9	98.2
Smiling	90.4	87.9	83.4	89.6	88.8	83.0	90.3	86.4	82.2	91.3
Eyeglasses	98.3	96.7	96.1	98.7	97.8	96.7	98.4	97.3	97.1	99.5
Wearing Hat	97.5	95.1	96.3	98.1	96.0	95.4	98.0	91.6	95.1	98.9
Wearing Lipstick	92.1	89.9	89.6	92.5	91.0	90.6	92.6	86.9	83.1	93.7
Wavy Hair	80.2	77.3	70.5	80.9	78.4	73.2	81.0	78.8	74.3	82.3
High Cheekbones	83.6	81.0	78.8	83.1	81.4	77.8	83.9	78.6	77.4	86.2
Mouth Slightly Open	88.0	85.4	83.7	88.4	87.0	83.3	87.7	83.8	83.6	90.7
Bald	95.9	93.1	93.4	96.0	93.7	94.9	96.0	90.2	92.2	98.9
Bangs	92.3	90.9	86.6	92.8	90.2	87.2	93.7	83.5	83.5	95.4
Heavy Makeup	86.3	83.2	81.1	87.2	84.0	82.1	87.0	82.8	79.8	89.9
Wearing Necktie	90.7	89.4	90.6	90.8	85.1	80.5	90.9	89.8	93.1	94.4
Blond Hair	91.3	87.1	86.9	92.0	89.5	87.4	91.6	87.2	86.4	95.1
Attractive	77.7	74.9	71.9	78.3	76.2	73.8	78.8	74.1	70.6	81.7
Young	83.4	75.6	72.4	84.4	81.3	80.4	83.0	69.0	66.8	88.3
Sideburns	92.5	89.0	89.1	93.1	91.6	89.3	93.2	89.7	88.2	97.5
Gray Hair	92.5	88.8	79.5	93.6	90.9	83.8	93.2	87.6	93.7	98.0
Black Hair	82.2	76.3	74.0	82.8	74.7	73.9	83.1	76.4	75.6	87.7
No Beard	90.2	87.7	88.5	90.9	89.4	89.0	92.1	84.0	85.7	95.7
Goatee	90.9	88.5	89.4	91.9	89.7	89.9	91.8	84.3	89.3	96.8
Arched Eyebrows	76.1	72.0	71.4	76.9	73.4	71.8	77.1	72.7	72.7	82.1
Mustache	89.7	86.6	85.9	90.2	88.2	87.1	90.5	85.2	79.2	96.6
5 o Clock Shadow	85.9	79.6	74.7	86.3	80.8	80.5	86.6	80.9	76.5	93.9
Big Lips	62.4	62.1	61.4	62.8	62.3	62.0	62.0	59.0	65.6	70.6
Bags Under Eyes	74.4	72.4	71.9	76.1	69.7	65.1	75.7	71.0	73.3	83.1
Receding Hairline	84.7	79.9	82.7	85.2	79.8	74.3	84.2	73.3	83.0	93.5
Pale Skin	86.4	82.9	76.1	85.4	84.1	74.5	86.1	86.7	85.4	95.8
Bushy Eyebrows	82.0	78.9	79.8	83.3	78.3	79.2	82.4	79.3	79.8	92.0
Rosy Cheeks	83.9	80.6	72.1	84.7	79.7	72.7	84.1	81.2	75.3	94.0
Double Chin	85.4	81.9	77.3	86.5	83.6	82.2	86.9	79.8	71.8	96.1
Wearing Earrings	74.9	67.5	61.3	75.8	65.4	59.0	77.6	72.1	64.9	85.6
Chubby	84.4	80.6	73.9	86.2	82.4	78.1	85.6	76.5	69.1	95.4
Straight Hair	70.6	59.6	53.3	73.4	61.5	58.0	72.9	57.7	48.4	81.9
Pointy Nose	65.3	62.6	65.3	66.6	63.6	62.0	66.2	65.7	66.5	76.6
Big Nose	72.6	69.6	68.7	74.0	70.5	69.5	72.0	68.5	62.8	83.9
Blurry	83.4	76.8	69.8	86.0	84.5	76.5	86.2	79.7	68.5	95.2
Brown Hair	74.0	67.3	64.2	74.6	64.3	55.9	73.7	74.2	70.1	86.2
Oval Face	60.4	59.4	60.3	61.6	60.5	60.2	63.8	59.0	62.2	73.7
Narrow Eyes	71.3	72.4	75.2	74.0	73.1	73.2	71.9	67.0	76.5	85.6
Wearing Necklace	62.7	54.8	54.7	64.3	58.8	53.1	66.7	61.9	49.6	86.2

Table 5.3: Average Attribute Accuracy

Method	Accuracy
LNets + ANets [23]	87.0
MOON [37]	90.9
AttCNN [13]	91.0
Multitask Baseline	90.2
Resnet50 (Layer)	83.1
Mobilenetv2 (Layer)	83.9
VGG16 (Layer)	83.9
Resnet50 (Channel)	79.5
Mobilenetv2 (Channel)	80.4
VGG16 (Channel)	78.6
Resnet50 (Neuron)	77.4
Mobilenetv2 (Neuron)	77.6
VGG16 (Neuron)	77.2

accurate values in comparison to the multitask baseline. Thus, even though the accuracy of these attributes decreases over the depth of network, the intermediate layers are still able to provide discriminative representations.

Table 5.3 shows results from additional baselines. Unfortunately, except for LNets + ANets [23] these methods do not provide the detailed accuracy of each attribute, and thus, it is not possible to obtain a fine-grained comparison with our results. Nonetheless, it shows that proposed multitask baseline achieves a comparable performance to the state of the art. Thus, it demonstrates that the multitask baseline is a reasonable representation of the accuracy of modern attribute classification. The average accuracy also allows for easier comparison between the layer, channel and neuron representation. On average, the best channel that was identified by the proposed analysis has an accuracy that is less than 4 percentage points bellow the best layer representation, with this gap increasing to 6 percentage points for the best neuron representation.

5.3 PLS and VIP Ablation Study

The previous analysis showed that there exists a few representational units are able discriminative some attributes almost as accurately as fully-supervised approaches. This analysis is extended in this section to also include low-scoring units, revealing how the attribute information is distributed inside the network. Since the VIP score is a measurement of relative importance of the neurons in the projection, scores obtained from two different projections are not comparable, and since a different projection is

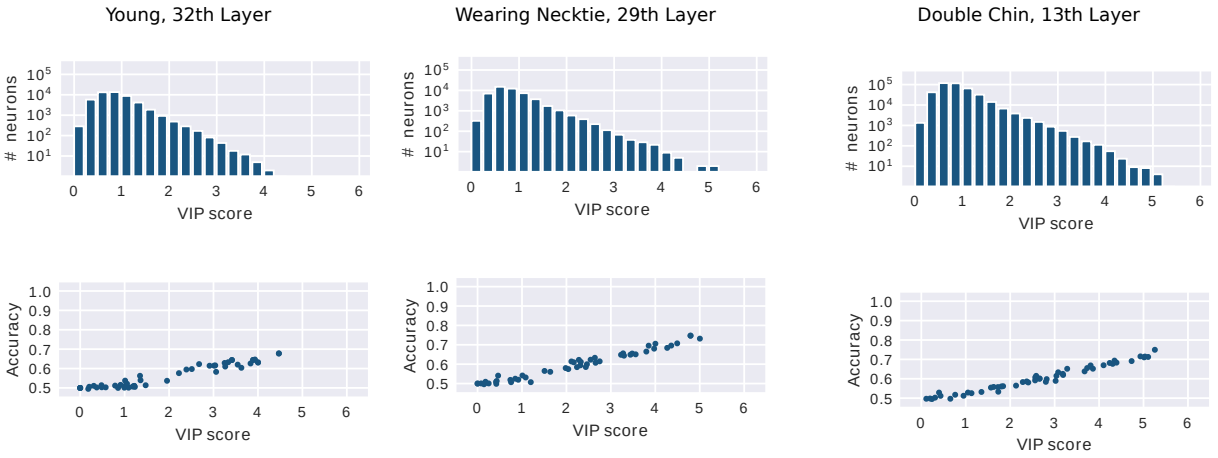

Figure 5.9: Top: VIP score distribution. Bottom: correlation between the score and accuracy for individual neurons. These plots illustrate the behaviour of three randomly selected models.

learned for each layer and attribute, it is not possible to aggregate results from multiple layers and attributes. Thus, instead of the mean result for all projections, Fig. 5.9 shows an analysis of the VIP score for three randomly selected PLS projections of the previous analysis.

The top row of Fig. 5.9 shows log-scaled histograms representing the distribution of VIP scores for all neurons in each of the selected layers. These results indicate that high-scoring neurons are very scarce, with the number of neurons decreasing exponentially for scores greater than one. To understand the relationship between the VIP score of a neuron and its discriminative power, we also analyse the accuracy of a classifier whose input is only a single neuron. More specifically, for each projection, 50 neurons are randomly selected, uniformly sampled according to their VIP score, and then a classifier is trained and validated for each individual neuron. The bottom row of Fig. 5.9 shows the performance of these classifiers according to the VIP score of the input neuron, revealing a strong linear relationship between these variables. In fact, by repeating this experiment for all layers and attributes, on average, a 0.93 correlation coefficient is obtained. For contrast, we also measure the correlation coefficient when the VIP score is replaced by the mean activation of the neuron for the positive samples, which is a popular metric to identify relevant units in visualization techniques. In this scenario, the average coefficient drops to 0.38, which is considered to be a weak correlation.

In summary, these results show that the VIP score is indeed an appropriate metric to determine which neural outputs are more relevant to specific attributes. Furthermore, the distribution of these neurons show there exists very few neurons that are highly discriminative of each attribute when evaluated individually, with the number of neurons decreasing exponentially for linear improvements of accuracy.

However, even though the obtained results demonstrated that most of the neurons in a layer have small VIP-score, and consequently, low predictive power for some face attribute, this result does not necessarily imply that they are not a crucial part of the representation. For instance, it could be the case that, in some layer, the representation of an attribute is more reliant on the combination of a large set of low scoring neurons from the previous layer rather than the equally discriminative small set of high scoring neurons. This hypothesis is evaluated through an experiment that consists of zeroing-out the outputs of the most relevant neurons of a specific layer, and then measuring how the accuracy of the classifiers of subsequent layers is affected. Table 5.4 shows the drops in performance of different representations that were caused by masking an increasing percentage of neurons from some layer. Each cell of the table shows three results, representing the decreased accuracy relative to masking a layer at the depths of $10, 25$ and 40, respectively.

Table 5.4: Accuracy decrease for subsequent representations of a layer at depth 10/25/40 with respect to the percentage of its top-scoring neurons zeroed-out.

	Representation		
Mask Size	Layer	Top-8 Channels	Top-8 Neurons
0.01	0.01/0.01/0.01	0.02/0.02/0.02	0.03/0.03/0.03
0.05	0.03/0.03/0.02	0.04/0.05/0.04	0.06/0.07/0.06
0.10	0.04/0.04/0.03	0.06/0.07/0.06	0.08/0.10/0.08

It should be noted that even the smallest evaluated mask size represents a far superior number of neurons than the top-8 that was previously used. Thus, the zeroed set of neurons are undoubtedly capable of accurately representing the attributes. However, removing these neurons from the representation of a layer does not seem to greatly affect the performance of the classifiers in subsequent layers. Thus, while our approach is able to find representative units across the dimensions of a layer, it seems that the combination of less representative units is more impactful for subsequent representations.

Chapter 6

Conclusions

The end-to-end optimization process of deep-learning networks produces representations that are both cryptic and discriminative. Consequently, the acceptance and improvement of these models are still limited by our inability to completely decode the meaning of the obtained representations. Previous attempts to understand deeply-learned features were mainly based on visualization techniques, which presented several limitations concerning the scalability and comparability of their results. This thesis proposed a quantitative approach to identify and compare the representation of several attributes encoded in the hidden layers of a deep face recognition network.

In essence, this approach consists of evaluating a set of classifiers to predict various attributes using features extracted from the intermediate layers of a deep network. This process allows direct and fair comparisons between several layers, channels and neurons by simply changing which representation is being extracted from the network. Even though the proposed analysis revealed some interesting patterns among all of the evaluated architectures, these findings may still not generalize for every face recognition CNN. In fact, the obtained network representations are likely to be influenced by several optimization parameters such as the training dataset, loss function, and data-augmentation procedure. Thus, one possible extension to this work is to measure the impact of these different design choices by gradually changing the training protocol and observing how each such modification affects the representation of the evaluated attributes.

The analysis in this work studied only binary facial attributes simply due to the convenience of isolating geometric transformations through facial keypoint alignment. Nonetheless, attributes from different domains could still be analysed with the proposed approach even if it is not possible to align their input images, though this would likely impair some representations. The impact of input alignment could also be studied in the face recognition domain by comparing the performance between classifiers with aligned inputs and classifiers whose inputs are subject to random geometric transformations. Even though the evaluated attributes were all binary encoded, the proposed approach can be easily adapted to both multi-class and regression problems through the replacement of the QDA classifier by another appropriate model.

The proposed analysis could also be expanded to include additional attributes.

Most of the attributes analysed in this thesis exhibited a very similar behaviour, with the majority of their accuracy improvements concentrated in first few initial layers of the networks. Discovering more attributes that are only well encoded in the deeper layers of the network may reveal what more interesting properties of deep networks. Unfortunately, determining these attributes may require an exhaustive search over many possible attributes, with each attribute requiring manual labeling. Furthermore, another important drawback of this expansion is that some attributes may encode sensitive information. Consequently, creating such a dataset raises ethical concerns as the data could be misused for mass identification of particular groups of people. Thus, the analysis of additional face attributes requires a careful consideration about the real-world impact of these attributes.

Finally, we also identify two potential applications of our obtained results in two different scenarios, namely, multitask learning and detection of adversarial attacks. While these scenarios were not explored in this work, we believe that they are potential directions for future works with the proposed method. In the case of multitask learning, the optimization of all different tasks is usually performed at the very last layer of the network, which may be sub-optimal if these tasks are not all closely related. In this scenario, the proposed analysis could be employed to identify the optimal depth to learn each sub-task. In the case of adversarial attacks, the application of our analysis comes from the observation that the representation of the attributes across the layers seems to be robust to manipulations of previous layers, at least in the case of zeroing-out some of the features. Thus, if it is possible to obtain a set of precise attribute classifiers for the real output classes of the network, then adversarial attacks may be detected if the attributes are not consistent with the final network output. For instance, in the case of face recognition, adversarial attacks could be detected if the identity label is not consistent with the attributes predicted across intermediate layers of the network.

Bibliography

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[4] Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[6] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–68, 2017.

[7] Prithviraj Dhar, Ankan Bansal, Carlos D Castillo, Joshua Gleason, P Jonathon Phillips, and Rama Chellappa. How are attributes expressed in face dcnns? *arXiv preprint arXiv:1910.05657*, 2019.

[8] M. Alves Diniz and W. Robson Schwartz. Face attributes as cues for deep face recognition understanding. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*, pages 566–572, Los Alamitos, CA, USA, may 2020. IEEE Computer Society.

[9] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016.

[10] Lennart Eriksson, Erik Johansson, N Kettaneh-Wold, Johan Trygg, C Wikström, and Svante Wold. *Multi-and megavariate data analysis*, volume 1. Umetrics Sweden, 2006.

[11] Claudio Ferrari, Stefano Berretti, and Alberto Del Bimbo. Discovering identity specific activation patterns in deep descriptors for template based face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE, 2019.

[12] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[13] Emily M Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

[16] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[17] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *International conference on biometric authentication*, pages 731–738. Springer, 2004.

[18] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[19] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[20] Been Kim, Justin Gilmer, Martin Wattenberg, and Fernanda Viégas. Tcav: Relative concept importance testing with linear concept activation vectors. 2018.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[24] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[25] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3):233–255, 2016.

[26] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62–69, 2012.

[27] Marvin Minsky and Seymour A Papert. *Perceptrons: An introduction to computational geometry*. 1969.

[28] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4467–4477, 2017.

[29] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*, pages 3387–3395, 2016.

[30] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*, 2016.

[31] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 55–76. Springer, 2019.

[32] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[33] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[34] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017.

[35] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[36] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer, 2005.

[37] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[38] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[40] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[42] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S Davis. Human detection using partial least squares analysis. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pages 24–31. IEEE, 2009.

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[44] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014.

[47] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[49] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[50] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[51] Herman Ole Andreas Wold. *Nonlinear estimation by iterative least square procedures*. 1968.

[52] Svante Wold, Michael Sjöström, and Lennart Eriksson. Partial least squares projections to latent structures (pls) in chemistry. *Encyclopedia of computational chemistry*, 3, 2002.

[53] Yudong Wu, Yichao Wu, Ruihao Gong, Yuanhao Lv, Ken Chen, Ding Liang, Xiaolin Hu, Xianglong Liu, and Junjie Yan. Rotation consistent margin loss for efficient low-bit face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6866–6876, 2020.

[54] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[55] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[56] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

[57] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017.

[58] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1136–1144, 2019.

[59] Yaoyao Zhong and Weihong Deng. Exploring features and attributes in deep face recognition using visualization techniques. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[60] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[61] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.