CONFIABILIDADE DE EXPLICAÇÕES VISUAIS DE MODELOS PROFUNDOS ATRAVÉS DE PERTURBAÇÃO ADVERSARIAL

DAN NASCIMENTO GOMES DO VALLE

CONFIABILIDADE DE EXPLICAÇÕES VISUAIS DE MODELOS PROFUNDOS ATRAVÉS DE PERTURBAÇÃO ADVERSARIAL

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

Orientador: Adriano Alonso Veloso

Belo Horizonte

Março de 2019

DAN NASCIMENTO GOMES DO VALLE

ASSESSING THE RELIABILITY OF VISUAL EXPLANATIONS OF DEEP MODELS THROUGH ADVERSARIAL PERTURBATION

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais – Departamento de Ciência da Computação in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Adriano Alonso Veloso

Belo Horizonte

March 2019

Valle, Dan Nascimento Gomes do

V181a Assessing the reliability of visual explanations of deep models through adversarial perturbation / Dan Nascimento Gomes do Valle. — Belo Horizonte, 2019
xxvi, 45 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Orientador: Adriano Alonso Veloso

1. Computação — Teses. 2. Aprendizado do computador. 3. Visão por computador. I. Orientador. II. Título.

CDU 519.6*82(043)

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

FOLHA DE APROVAÇÃO

Assessing the Reliability of Visual Explanations of Deep Models through Adversarial Perturbation

DAN NASCIMENTO GOMES DO VALLE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Adriano Alonso Veloso - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Nivio Ziviani
Departamento de Ciência da Computação - UFMG

Prof. William Robson Schwartz
Departamento de Ciência da Computação - UFMG

Prof. Eduardo Alves do Valle Junior
Departamento de Engenharia de Computação e Automação Industrial - UNICAMP

Belo Horizonte, 27 de Março de 2019.

To my family and friends, who have always supported me.

Acknowledgments

I would first like to thank my family, who supported and motivated me at all times. Also my friends, who have always been with me in every step of my masters. All the conversations we had, made me evolve quickly. Special thanks to the professors Adriano Veloso and Nivio Ziviani, who served as inspiration and guided me. Our meetings, full of advice and teaching, were responsible for how far I got.

"You cannot teach a man anything; you can only help him to find it within himself."
(Galileo)

Resumo

O crescente interesse em redes neurais profundas complexas para novas aplicações exige transparência em suas decisões, o que leva a uma necessidade de explicações confiáveis das decisões tomadas por esses modelos. Trabalhos recentes propuseram novos métodos de explicação para apresentar visualizações interpretáveis da relevância das instâncias de entrada. Esses métodos calculam mapas de relevância que geralmente se concentram em diferentes regiões de pixel e são comumente comparados por inspeção visual. Isso significa que as avaliações são baseadas na expectativa humana, em vez da real importância das *features*. Neste trabalho, propomos uma métrica eficaz para avaliar a confiabilidade da explicação de modelos. Essa métrica é baseada nas mudanças da resposta da rede, resultante da perturbação das imagens de entrada de maneira adversarial. Essas perturbações consideram todos os valores de relevância e suas inversões (irrelevância), de modo que a métrica tenha características de precisão e revocação. Também propomos uma aplicação direta dessa métrica para filtrar mapas de relevância, a fim de torná-los mais interpretáveis sem a perda de explicações importantes.

Nós apresentamos uma comparação entre alguns métodos de explicação amplamente conhecidos e seus resultados pela métrica proposta. Também expandimos os resultados para uma discussão sobre técnicas de visualização e a quantidade de informação que é perdida para torná-las mais interpretáveis e intuitivas. Em seguida, mostramos os resultados do nosso método de filtragem que aborda esse problema. Além disso, apresentamos uma análise aprofundada das propriedades da métrica que a tornam apropriada para uma variedade de tarefas. Nela, observamos a importância de usar a irrelevância, a robustez a valores aleatórios e imagens classificadas incorretamente, e a correlação entre a métrica e a perda do modelo avaliado.

Palavras-chave: Visão Computacional, Aprendizado Profundo, Aprendizado de Máquina Explicável.

Abstract

The increasing interest in complex deep neural networks for new applications demands transparency in their decisions, which leads to the need for reliable explanations of such models. Recent works have proposed new explanation methods to present interpretable visualizations of the relevance of input instances. These methods calculate relevance maps which often focus on different pixel regions and are commonly compared by visual inspection. This means that evaluations are based on human expectation instead of actual feature importance. In this work, we propose an effective metric for evaluating the reliability of the explanation of models. This metric is based on changes in the network's outcome resulted from the perturbation of input images in an adversarial way. These perturbations consider every relevance value and its inversion (irrelevance) so that the metric has characteristics of precision and recall. We also propose a direct application of this metric to filter relevance maps in order to create more interpretable images without any loss in essential explanation.

We present a comparison between some widely-known explanation methods and their results using the proposed metric. We also expand the results into a discussion on visualization techniques and the amount of information lost to make them more interpretable. Then, we show the results of our filtering method which tackles this problem. In addition, we further present an in-depth analysis of the properties of the metric which make it appropriate for a variety of tasks. It shows the importance of using the irrelevance, the robustness to random values and misclassified images, and the correlation between the metric and the loss of the model evaluated.

Palavras-chave: Computer Vision, Deep Learning, Explainable Machine Learning.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The success of deep neural networks to model increasingly difficult problems brought the attention of several areas to these techniques, such as healthcare [Miotto et al., 2018; Shen and Suk, 2017] and economics [Hartford et al., 2017]. To achieve these results, the models are constantly evolving and becoming ever more complex. This increase in assertiveness, though, usually comes at the expense of interpretability. Referred to as black-boxes, users often avoid such systems in practical applications due to the lack of explanation in their decisions. This lack of transparency brings distrust to them, who frequently prefer to use simpler interpretable methods.

The interpretation of simpler models such as decision trees and classification rules are more straightforward as they explicitly show each consideration made until it reaches a final decision [Freitas, 2013]. However, the works in these areas commonly propose deep neural networks as an appropriate solution [Liu et al., 2014; Payan and Montana, 2015]. These complex models are harder to understand and require more advanced methods.

In some applications, the reasons considered for a decision are as important as the decision itself. When using machine learning to diagnose Alzheimer's disease, for example, just claiming its occurrence is not enough and can be restricted even by regulations [Goodman and Flaxman, 2017]. It requires further explanation of what patterns or region of the brain led to that decision as these predictions have serious consequences on patients' lives. Therefore, reliable explanations are essential in such problems. This reliability in visualizations represents the correct relevance of each feature the model considers in a classification. Then, it prevents these models to expose wrong visualizations in which an observer may accept as correct just because of the presence of expected patterns.

There are several techniques which tackle the problem of finding relevant infor-

mation that can present the reasons for a decision [Simonyan et al., 2013; Ribeiro et al., 2016; Baehrens et al., 2010; Fong and Vedaldi, 2017]. Both the developer of the model and the users benefit from understanding this importance given to the features. This is because unwanted patterns learned by the model, such as biases, can be revealed while using it to debug the model, which later leads to enhancements and improved methods [Cadamuro et al., 2016; Lakkaraju et al., 2017].

The models trained on unstructured data, such as multimedia and text, are known to be more complex to extract useful interpretations than on structured data. This is because, in most cases, they do not have specific features that are relevant to a whole dataset. Instead, each instance has its own individual features importance. For example, when detecting dogs in images, the positions they appear in the image can change even when the appearances of the dogs are very similar. Therefore, we will have the same patterns in different regions of images. This work focuses on the interpretability of these types of problems.

The analysis of the patterns learned by deep neural networks is a topic recently addressed [Zeiler and Fergus, 2014; Erhan et al., 2009; Le, 2013; Montavon et al., 2011]. Several works are focused on understanding specific patterns learned by neurons [Erhan et al., 2009; Le, 2013; Yosinski et al., 2015]. For this, they maximize activation functions using subsets of data and observe the resulting shapes. Other techniques use the model's parameters in order to expand their representations [Dosovitskiy and Brox, 2016; Mahendran and Vedaldi, 2015]. More recent techniques which are important to the problem discussed in this work focus on understanding the decisions for individual instances in a model on pixel-level [Ribeiro et al., 2016]. In this case, each prediction is linked to the importance of each pixel in the input image. There are several approaches that find this interpretation, which goes from saliency maps [Simonyan et al., 2013; Smilkov et al., 2017] to contributions from the decomposition of predictions [Bach et al., 2015; Binder et al., 2016].

Although there are approaches that find different types of interpretations for decisions of trained models, it is important to understand the main objectives of each one and how to use them in the right way. We must consider that deep neural networks have special behaviors and can be easily confused [Nguyen et al., 2015; Goodfellow et al., 2014; Szegedy et al., 2013] . Therefore, techniques that explain these models must aim to present the real importance of the features in the decision, following the characteristics of the models. To do this, we should not consider visual interpretations that make sense to a person who is observing them as the same used by the model in a prediction. While visualizations that seem to be expected by an observer can have several advantages in an application, they can be misleading and different from the

real importance given to features in a trained network.

Adebayo et al. [2018] presents a series of experiments which support the idea that simple visual inspection is not the proper way to evaluate the relevance maps generated. In their results, some of the most used techniques create relevance maps which are independent of the prediction. This means that the explanation is based solely on the input image. Therefore, it is not an explanation related to the model. Our work uses this conclusion as a basis and aims to measure how much a relevance map created for a model and an input image is actually reliable with respect to how much the calculated values influence a classification.

1.1 Thesis Statement

The explanation of the predictions of Deep Neural Networks is still a complex unsolved task, resulting in a variety of explanation methods which present different relevance visualizations, often compared by simple visual inspection. The main hypothesis of this thesis is that the relevance maps from such methods are algorithmically comparable. Thereby, the most accurate one represents the actual importance of the features of the evaluated model. The aim of this thesis is to show an effective use of all relevance values to perturb the input instances in an adversarial way by aiming to minimize and maximize their magnitude. It leads to measurable changes in the outcome of the model which results in a metric of reliability.

1.2 Our Solution

We designed our solution around the reaction of models to when they undergo changes in input images on the directions presented by the gradients, as shown in [Goodfellow et al., 2014; Szegedy et al., 2013]. They argue that minimal changes to an image, even if not noticeable to humans, can completely change the decision of a model. This shows that changes to low-level features guided by their importance strongly influence the values presented in higher layers after being calculated. Thus, we exploited this along with the map of visual relevance of the explanation techniques to measure the correct direction which is sufficient to create maximum or minimum changes in higher layers.

The studies on the Selectivity Property [Bach et al., 2015; Samek et al., 2017] assert that a model must agree with its explanation. Thus, features in the input image with higher respective relevance score in an ideal relevance map are the ones which

most influence the output of the model classification. In order to maximize alterations in the outcome of a model by perturbing the input image, we retain the sign values of the gradient to use it along with the relevance images.

Then, we make perturbations similarly to an ascending step directly on the image with the values based on the relevance in the positive direction of the gradient, approaching a maximization in the error with minimal input modification. By measuring the influence of the relevance maps in the images, it is possible to define a metric which compares the reliability of the visual explanation from the methods for a given model and an input image. Our proposed metric is thus called Adversarial Perturbation Explanation Metric (APEM).

It is also important to consider the low scores of relevance to avoid privileging methods whose maps restrict relevant features to a few concentrated high values. Besides the perturbation from relevance scores, we introduce the same approach to their opposite values, which we call irrelevance. A correct relevance map implies accurate irrelevance values which require higher perturbation by the Selectivity Property. This balances the precision and the recall of the resulting maps, which all relevant values have to be detected and also be correct.

The particularities of the APEM allow its use as an approach to a variety of other problems. For instance, we propose a filtering algorithm which makes explanation methods more interpretable to an observer while keeping all the essential information in the relevance maps. We use APEM scores for the images as constraints and the values calculated for the modified maps cannot be lower than the previous ones. This results in images with minimum noise while ensuring no loss in explanation.

1.3 Contributions

The main contribution of this work is to measure the reliability of methods that explain decisions of models. The proposed metric exploits the behavior of deep networks when presenting adversarial examples. We state the following claims and contributions:

- We compare methods that produce visually relevant interpretations of models and how they actually affect deep networks outputs. Moreover, we debate what each work proposes. When using their relevance as a guide to perturb the input images, we show how models are tricked to change their answers.

- We introduce a metric that can relatively show how different techniques, which create images that look convincing to the human's knowledge of the world, are

trading off actual relevance to the model against more understandable shapes. We also show how visualization techniques can abdicate important information in order to make images more comprehensible. This metric substitutes visual inspection when evaluating explanation metrics. It establishes the use of relevance values along with the irrelevance, and we present several experiments of its properties.

- We present a new approach based on the proposed metric to extract relevance from images which shows visually understandable shapes and is also relevant to models in their predictions. It is used as a filter to relevance maps and can be applied to the results of any explainable method so it becomes more interpretable without any loss of essential information.

1.4 Organization

The rest of this dissertation is structured as follows. In Chapter 2, we present the works on understanding neural networks and the explanation metrics. Then, we show their vulnerability, some unwanted characteristics of the explanation methods and the related works on evaluating these methods. In Chapter 3, we describe our proposed metric and how it can be used to filter relevance maps, creating less noisy visualizations. In Chapter 4, we describe the dataset and the methods that are compared by the metric. Then, we present the experimental results and some points on the lost of explainability in visualization techniques. After that, we show some properties of the APEM in other proposed experiments. Finally, in Chapter 6, we present the concluding remarks and future work.

Chapter 2

Background and Related Work

In this Chapter, we present an overview of the works that aim to understand neural networks and the efforts on explaining their outcomes. Then, we introduce the studies on the properties of them and how deep visualizations can be unreliable in some cases. Finally, we show the approaches to evaluate the quality of the relevance maps from explanation methods.

These relevance maps come from visualization methods that calculate relevance values related to each pixel of the input. So, it creates images with the same dimensions of the input in which higher values are given to the ones with more importance in the classification by the analyzed trained model.

2.1 Understanding and Explaining Neural Networks

The rapid progress of Deep Learning techniques made it so focused on achieving better results in a variety of tasks, such as object detection in images [Girshick et al., 2014; Girshick, 2015; Ren et al., 2015], that methods to explain models did not keep up at the same pace. The non-linearity present in these models makes it difficult to really understand how they are learning and the considerations in each decision.

Some authors worked on understanding the neurons in complex systems of deep neural networks by visualizing interpretations of higher level features in Autoencoders [Erhan et al., 2009; Le, 2013] and Deep Belief Networks [Erhan et al., 2009]. These studies show how models can find patterns that are similar to what human consider relevant in the domains analyzed. Montavon et al. [2011] approaches this problem of what is represented in the models' units by examining the complexity of each layer. They concluded that higher layers represent the input in simpler and more accurate forms. All these efforts on interpretable visualizations of features to understand neu-

ral networks culminated in tools that easily displays information about trained layers [Yosinski et al., 2015], so it could be widely studied.

Understanding what is responsible for the performance of deep models, especially in Convolutional Neural Networks (CNNs), is essential to improve the current models and find situations in which they fail to infer correctly. Zeiler and Fergus [2014] proposes Deconvolutional Networks that map from a higher layer back to the pixels in the input space in an already trained model. Therefore, for each image presented, the authors create a visualization that shows the patterns in the input which resulted in a specific activation in further layers. Other techniques tackle this problem by inverting the input images representations and analyzing the information which is maintained in them [Mahendran and Vedaldi, 2015; Dosovitskiy and Brox, 2016]. As a result, many ideas on the properties of the feature representations of deep models arise.

Simonyan et al. [2013] presents a gradient-based approach, comparable to the Deconvolutional Networks, to visualize the patterns in the input image. The proposed saliency maps use the gradient of the calculated classes scores. Moreover, this technique generates images which represent the notion of the classes learned by the model when maximizing these scores. Smilkov et al. [2017] makes use of the saliency maps but tackles the problem of the amount of visual noise produced. In order to do so, they create a sample of noisy copies from an input image and average the maps calculated for them. Other works also use the gradient in different ways to achieve contrasting visualization results [Springenberg et al., 2014; Selvaraju et al., 2016; Sundararajan et al., 2017]. Ribeiro et al. [2016], on the other hand, explains predictions by learning an interpretable model locally around them. The authors apply it in a variety of classification tasks for image and, also, text.

Another group of related works which aims to understand non-linear predictors are focused on calculating scores of relevance for pixels by applying a technique which redistributes the relevance from the output back to the input image assuring a conservation property layer-wise [Bach et al., 2015; Binder et al., 2016]. Consequently, neurons that contribute more to the others in the following layers, have higher scores. In [Lapuschkin et al., 2016], this method is extended to Fisher Vectors and serves as a tool to show how differently models consider certain regions of images as relevant or not. In addition, their conclusions come with questions about the reliability of models' classification which can be, for example, biased by context.

2.2 Reliability of Deep Visualizations

Deep neural networks are complex systems and a variety of flaws are detected in CNNs and explored as properties [Nguyen et al., 2015; Goodfellow et al., 2014; Szegedy et al., 2013]. The works in [Szegedy et al., 2013] and [Goodfellow et al., 2014] investigate the counter-intuitive property called Adversarial Examples, which we further explore in this dissertation. They show how susceptible these models are to directed perturbations imperceptible to humans, which make them misclassify. Furthermore, they present the linear nature of neural networks as the main cause of such vulnerability.

Nguyen et al. [2015] goes deeper into the subject and questions the differences actually considered between patterns seen by humans and deep models by creating images that are simply noise and have extremely high confidence. In our work, we follow this doubt and ask if visualization methods that are applied in such networks and considered better than others because they represent patterns that are expected by humans are really expressing the importance of the features to the model itself. To properly debug and understand a model, the method of explanation has to show the true relevance of pixels in an interpretable manner even if noisy and not presented with an appealing appearance for those who see.

Adebayo et al. [2018] works exactly on this question and proposes a methodology based on randomization tests in order to evaluate the reliability of explanation methods. They present two tests which check some characteristics of the explanation approaches, excluding them of some tasks which require these characteristics, such as debugging models and revealing unintended effects learned by the model, and are divided into:

- Model Parameter Randomization Test: Compares the relevance maps of a trained model with the ones of a randomly initialized untrained network of the same architecture. If the maps depend on the learned parameters of the model, the results of the two cases should differ significantly. Otherwise, they are insensitive to the parameters of the model. The authors subdivided this test into two forms:
  - Cascading Randomization: randomize the weights of a model starting from the top layer, successively, all the way to the bottom layer.
  - Independent Randomization: independent layer-by-layer randomization with the goal of isolating the dependence of the explanations by layer.
- Data Randomization Test: Compares the relevance maps of a model trained on a labeled dataset with the ones of a model with the same architecture but trained on a copy of the dataset in which they randomly permuted the labels. If the maps

depend on the labeling of the data, the results of the two cases should differ again. Otherwise, they are insensitive to the relationship between the input images and the original labels.

Explanation methods insensitive to randomizing labels are not able to explain cases in which there is a relationship between the images and labels present in the data generating process. Furthermore, if insensitive to both cases, the approach does not present reliable information on tasks which depend on the model parameters or the relationships between inputs and outputs.

They provide experiments on several methods and analyze the results comparing them to a simple edge detection algorithm, which does not require a training process and is dependent only on the image. Moreover, they find widely used methods which are independent in both tests and conclude that visual inspection is a poor way of evaluating explanation results. These results imply the need for techniques which compare methods not simply by visualization, but based on a more effective approach to show the actual importance of the features to the outcomes of the models.

2.3 Evaluation of Explanation Methods

Comparisons between methods of deep visualization are commonly qualitative and compare selected examples of maps of relevance which restrict their values to the pixels that humans focus the most. Samek et al. [2017] tackles the problem proposed here, providing a quantitative metric that evaluates how the relevant areas affect the correct prediction of a given model, and is the closest work to ours. Their approach consists of sequential random perturbations in the sorted relevant regions of the images and evaluates, for each, when the model starts to misclassify.

While they consider block regions of relevance and apply random changes on it, we propose that the relevances calculated to a specific class can guide hardly noticeable perturbations in the whole image and effectively cause failures in the model when the relevance scores are really representing the importance of the features to the model.

Montavon et al. [2018] discuss the desirable properties of explanations and possible evaluation metrics by using two of them. They argue that it is important to define the characteristics of a good explanation at a more abstract level. Then, they support a quantitative assessment of the Continuity and Selectivity properties, as explained below.

2.3.1 Explanation Continuity

The property of continuity ensures that if two data points are nearly equivalent, then the explanations of their predictions should also be nearly equivalent. They show that continuity can be quantified by seeking the greatest variation of the explanation $R(x)$ in the input domain:

$$\max_{x \neq x'} \frac{||R(x) - R(x')||_1}{||x - x'||_2}$$

2.3.2 Explanation Selectivity

The explanation redistributes relevance to variables that have the strongest impact on the classification. Then, the property of selectivity states that removing features with attributed relevance should reduce evidence at the output. The works in [Bach et al., 2015; Samek et al., 2017] quantify this property by measuring how fast an evaluated function starts to decrease when removing features with the highest relevance scores as in Algorithm 1.

Algorithm 1 Explanation Selectivity

1: **repeat**
2: **record** current function value $f(x)$
3: **find** feature i with highest relevance $R_i(x)$
4: **remove** feature i ($x \leftarrow x - \{x_i\}$)
5: **until** all features have been removed

Chapter 3

Adversarial Perturbation Explanation Metric

We tackle the quantitative evaluation of deep visualizations task by using an approach based on guided perturbations to the original images until the model loses the ability to correctly classify them. The existing visualization methods generate one relevance value for each pixel, resulting in images with the same dimensions in width and height, in which the unit values are the relevances $R = [[r_{i,j}|0 \leq r_{i,j} \leq 1]]$. Thus, $r_{i,j}$ values where $r_{i,j} \simeq 1$ means that they have higher importance in the classification by the trained model used to create R, and $r_{i,j} \simeq 0$ means the opposite.

Each method uses a different framework to create the relevances and they compare their results by presenting qualitative results based on intuition and the clarity of what is visible in their maps. We propose in our work (*i*) to compare these visualizations as a matter of importance to the model, and (*ii*) to use the same line of work to create simpler and more understandable visualization that can be used on the visualizations to make it more understandable while keeping their correct importance. We are also going to discuss the trade-off of creating relevances that are more appealing but lose explainability, and compare to our approach that retains as much information as possible.

3.1 Comparing Relevances

Once a model is pre-trained, the relevance images can be extracted from a set of input images and used in the evaluation. Therefore, the quality of the explanations depends on the classification ability of the model, since they are calculated based on the learned parameters and its prediction capacity. Our method, however, does not create a general

measure that can be used to compare different models with diverse capacities. It ranks the explanation for a given network, giving higher priorities to the ones that most depicted the information used in the classification.

Analyzing the extracted relevances, we assume that the higher values are the most important to the correct prediction and their absence would result in the greatest impact in the output. The lack of these features would be enough to make the model misclassify. On the other hand, removing the given irrelevant features should result in less impact on the model. It does not prevent the model from misclassification, but the required perturbation on these features are higher when compared to the previous. Previous works have applied perturbations in the image regions randomly [Samek et al., 2017] but, here, we propose to change the whole input image so we consider relevance pixel-wise. These perturbations are hardly noticeable and follow the studies of directed changes on pixels to harness networks in [Goodfellow et al., 2014].

First of all, we recall that smaller changes in high relevances should deviate the decision, and these values are the ones that most influenced the outcome of the model. Additionally, they are positive values that do not have a sense of direction in which they should follow to effectively trick the model. To approach this problem, we propose a Directed Relevance R_{dir}, which attributes the direction of the relevances by using the sign of the gradients given by the model when predicting the same class the relevances are representing. To make it fair among the methods that vary the total relevance distributed in the image, we normalize the values in R by the l_1-norm. Then, for a network with parameters θ, an input image x and the target y associated to x, we use the sign of the gradients from a loss function $J(\theta, x, y)$ to direct the relevances, as shown in:

$$R_{norm} = \frac{R}{norm(R)}$$

$$R_{dir} = R_{norm} \odot sign(\nabla_x J(\theta, x, y))$$

Accordingly, the R_{dir} which correctly represents the model would need the minimum perturbation to maximize the deviation from the correct answer, as it is correlated to one gradient ascent step in the pixels instead of in θ. Therefore, there is a minimum ϵ^- value which makes the model change its prediction for the perturbed image x' created from x as seen in:

$$x' = x + R_{dir} \times \epsilon^-$$

Expanding this thought, we extract the values of irrelevance from R by making $R^- = 1 - R$. Moreover, the best R^-_{dir} calculated from R^- is the one the takes longer to create doubts in the predictions, as changes in pixels which are not important do

not cause significant changes in the output. This results in a maximum ϵ^+ value and, consequently, a gap between it and ϵ^-.

The Adversarial Perturbation Explanation Metric (APEM) proposed here is the average of all the gaps for a given model and the explanations extracted from a set of images. Given this, our metric has the purpose of comparing which applied explanation method most represents the importance of the features for the same model and set. Then, it is not a global metric for the methods, but relative to the exact analyzed situation. We can express the APEM for n instances as:

$$APEM = \frac{\sum_{i=1}^{n} (\epsilon_i^+ - \epsilon_i^-)}{n}$$

3.2 Making Relevance Interpretable

Since we can now measure the reliability of a relevance map, the APEM presents the amount of useful explanation lost when simplifying the relevance map to make it more interpretable. A good simplification is the one which does not reduce the APEM value of the original map. The pixels which are less relevant to a prediction have smaller values in the map and, in most cases, no ability to cause changes in the outcome of a prediction when not considered as relevant to the APEM. By zeroing these relevance values, their irrelevance values become one and the APEM score changes to the new map. If the metric remains unchanged, the pixels are considered to have no influence in the prediction by themselves and can be excluded from the visualization so it is more understandable and still holds the same explanation.

We propose here an algorithm to filter relevance maps, removing the noise in it so humans can easily interpret. It zeroes the least relevant non-zero values iteratively until any change in the relevance map causes a reduction in the metric, as seen in Algorithm 2.

This method is applicable in any relevance map calculated for an image and a trained model. While other methods have to deal with the trade-off of losing information to make deep visualizations more interpretable, the one presented here works as a way to enhance interpretation while keeping the same explanation of features considered by the model. Therefore, we have more reliable visualizations which are also understandable by users of some application using it.

Algorithm 2 Explanation Filter

```
function explanationFilter(relevance_map, image)
    apem ← APEM(relevance_map, image)
    repeat
        best_relevance_map ← relevance_map
        best_apem ← apem
        min_val ← get_minimum_positive_value(best_relevance_map)
        relevance_map ← zero_where_equal(best_relevance_map, min_val)
        apem ← APEM(relevance_map, image)
    until apem < best_apem
    return best_relevance_map
```

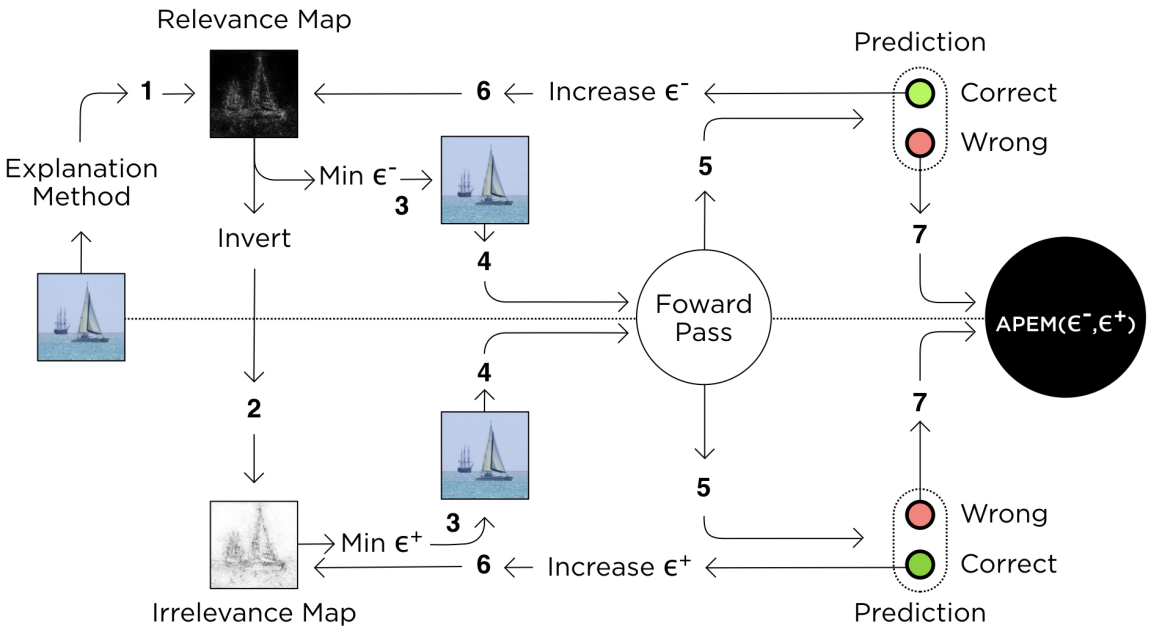


Figure 3.1: Diagram of the APEM flow for an input image. The transitions are enumerated sequentially from the first step to the last and the ϵ values are stored and modified throughout the process.

3.3 Overview

This section presents an overview of the proposed metric. We also show a step-by-step example following the stages of the diagram in Figure 3.1. The diagram divides the process into seven main steps, which are:

1. The evaluated explanation method calculates the relevance map R from the original image.

2. The proposed irrelevance map R^- is then created by inverting the values of the

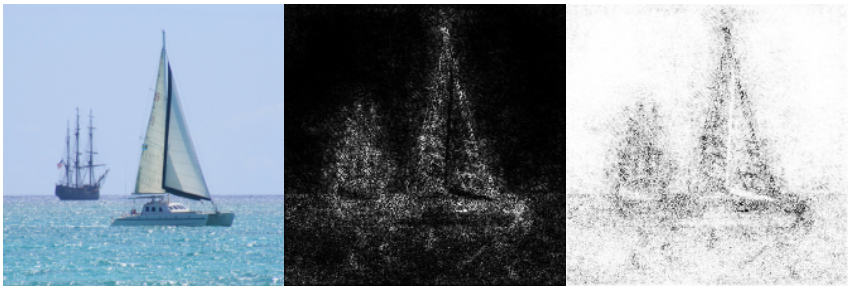


Figure 3.2: Example of the relevance map (middle) and its irrelevance map (right) for a sampled input image (left).

first map using the formulation $R^{-} = 1 - R$.

3. Both the relevance and irrelevance maps are used to perturb the original image in the positive direction of the gradient, controlled by a minimum perturbation value ϵ, creating their respective input images to the model. A human observer hardly notices the difference between these perturbed images and the original ones.

4. The model receives the perturbed image as input.

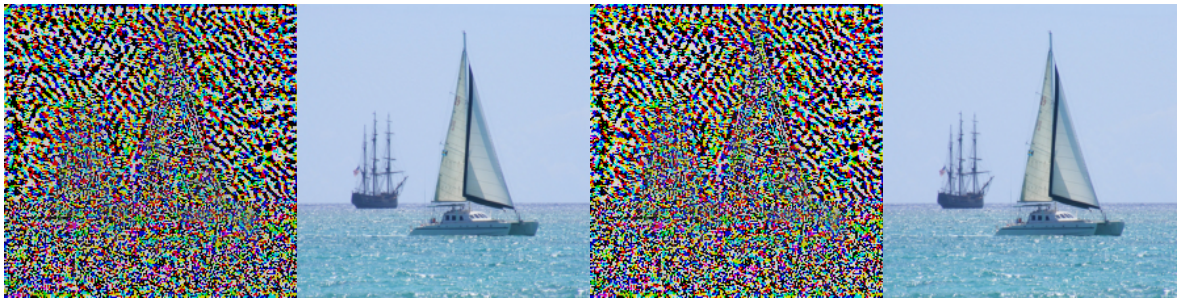
5. The model outputs a prediction, which can be correct when compared to the ground truth or a misclassification. A correct prediction means that the minimum ϵ in step 3 is not great enough to make the model misclassify while the other case means that the input image caused a change in the original decision of the model.

6. Every correct classification makes the value of ϵ increase so new perturbed images can be calculated and used to make the model misclassify.

7. Wrong predictions define the minimum correct value for ϵ which sets the metric score.

As an example, Figure 3.2 shows the relevance and irrelevance maps for a sampled image. The gradient was used as the method to create the relevance image and the irrelevance calculated from it. Then, the next steps consist of perturbing the original image by using the maps presented in the figure until the model changes its output. Figure 3.3 concludes the example showing each applied perturbation and the resulting images for the relevance (3.3a) and irrelevance (3.3b) maps presented and their respective ϵ values until they become the minimum values which affect the outcome of the model. Even though the perturbation is minimal and the resulting images look the same compared to the original, the model misclassifies them with low ϵ values.

(a) Perturbation and resulting image from the relevance map with $\epsilon = 1$ and $\epsilon = 2$ from the left to the right. The model has a correct prediction when $\epsilon = 1$ and misclassifies when $\epsilon = 2$.



(b) Perturbation and resulting image from the irrelevance map with $\epsilon = 3$ and $\epsilon = 4$ from the left to the right. The model has a correct prediction when $\epsilon = 3$ and misclassifies when $\epsilon = 4$.

Figure 3.3: Example of the process iteration for a sampled image. It presents the applied perturbation values and the resulting perturbed images which are forwarded through the model in the relevance (a) and irrelevance (b) steps until the model misclassifies (perturbation images shown here are normalized by scaling between 0 and 1 to make it easier to see). This example has a final value of $\epsilon^- = 2$ and $\epsilon^+ = 4$.

Chapter 4

Experiments

In this chapter, we present the techniques applied to extract the relevance from the datasets and their results on the APEM. We also show how more interpretable visualizations may affect the real explanation of the features and how our technique can effectively tackle this problem, creating more understandable images while keeping the same APEM value of the gradient.

4.1 Dataset and Framework

The relevance images were calculated from the ImageNet Large Scale Visual Recognition Challenge 2012 dataset (*ILSVRC2012*) [Russakovsky et al., 2015]. It is a competition whose goal is to estimate the content of photographs for the purpose of retrieval and automatic annotation using a subset of the large hand-labeled ImageNet dataset as training. The challenge aims to identify the main objects in test images labeled especially for this competition and are not part of the previously published ImageNet dataset. The role of the algorithms is to produce the respective labels of the objects in these images.

The training data from the ImageNet contains 1.2 million images and 1000 categories. The validation and test data for this competition consist of 150,000 photographs without duplicates. They were collected from search engines, hand-labeled with the 1000 categories of objects, and are not among the ImageNet training data. In addition, the validation set contains a random subset of 50,000 of these labeled images, and the remaining images are in the evaluation test set.

We used sets of 5,000 random images from the validation set to create the relevances for each explanation method evaluated, one set of correctly classified images and another of misclassified ones. The methods were implemented using PyTorch [Paszke

et al., 2017], which provides a VGG-16 model [Simonyan and Zisserman, 2014] trained on the *ILSVRC2012* training set.

4.2 Explanation Methods

The visualization function is standardized for all explanation methods. For each relevance matrix, we apply the modulus of the values and clamp them to an upper-bound of the ninety-ninth percentile of the resulting absolute values. Then, we multiply the new values by their respective original pixel values, resulting in a cleaner visualization as proposed in [Smilkov et al., 2017]. By doing this multiplication, we go back to the debate of how much alteration in the explanation we make when simplifying the relevance map to make it more appealing. This stage brings shapes even easier to understand to the map. However, it can also create side effects, such as false relevance values as a result of the multiplication itself, which we thoroughly analyzed in Section 4.4. As we are using RGB images, we reduce the final number of channels to just one by summing the three values along the dimensions, and we normalize it to the range $(0, 1)$.

Besides the use of the gradient directly interpreted as a relevance image, we compared two distinct techniques: Smooth Grad and Layer-wise Relevance Propagation (LRP). We further explained them below.

4.2.1 Smooth Grad

Smooth Grad [Smilkov et al., 2017], like other techniques based on interpreting the gradient as a sensitivity map, uses the gradient of a class score function with respect to an input image to explain the output of deep networks. For the image classifier in our experiments, it shows the importance of the pixels to the calculated outcome. Mainly, the technique averages the sensitivity maps generated to a sample of images created by adding noise to a selected initial image. As its objective is to reduce noise in the deep visualization, it is applicable to other methods which also create sensitivity maps.

More specifically, for a given input image x and a model which classifies an image into $class(x)$ by calculating the highest score of an activation function S_c, for a class set C where each class $c \in C$. It can be expressed as

$$class(x) = argmax_{c \in C}(S_c(x))$$

Then, for each image x and the proposed model, the sensitivity map $M_c(x)$ can

be calculated from the result of the differentiation M_c with respect to the input image x. Representing the gradient of S_c as ∂S_c, the sensitivity map can be defined as

$$M_c(x) = \frac{\partial S_c(x)}{\partial x}$$

Smooth Grad takes into consideration the characteristics of the map M_c to calculate important regions and proposes the use of a smoothing of the gradient ∂S_c, by applying a Gaussian kernel, instead of the raw gradient ∂S_c. As a result, the algorithm averages the sensitivity maps of the random samples calculated in a neighborhood of x, formulated as

$$\hat{M}_c(x) = \frac{1}{n}\sum_{1}^{n} M_c(x + \mathcal{N}(0, \sigma^2))$$

where n is the number of samples to average over, and $\mathcal{N}(0, \sigma^2)$ the Gaussian perturbation with standard deviation σ. The variation of the hyperparameters controls:

- σ: balance the sharpness of the sensitivity map and maintain the structure of the original image
- n: make the estimated gradient smoother as n increases

4.2.2 Layer-wise Relevance Propagation

The Layer-wise Relevance Propagation (LRP) [Binder et al., 2016] computes the relevance of the pixels of a selected image by considering their impact on the output of the model. To create a framework for the calculation of the contribution values in the input image, this method takes into consideration the characteristics of the general function of neurons in the feed-forward deep networks, expressed as

$$x_j^{(l+1)} = g\left(0, \sum_i x_i^{(l)} w_{ij}^{(l,l+1)} + b_j^{(l+1)}\right), \text{ e.g. } g(z) = max(0, z)$$

where x are values of the neurons, w and b the weights and the biases connected to the neurons respectively, and where i represents the index in a layer (l) which connects to a next layer $(l+1)$ indexed by j and forwards their values. The model learns the parameters w and b in the training phase.

LRP proposes the use of the graph structure presented with trained parameters to redistribute the relevance value at the output of the network back to the pixels. The relevance is propagated until it reaches the input, generating the pixel scores $R_p^{(1)}$, by

applying in each layer the redistribution rule

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)} \text{ with } z_{ij} = x_i^{(l)} w_{ij}^{(l,l+1)}$$

where each neuron relevance value R_i in layer l is calculated based on its contribution to all neurons in the next layer $(l+1)$. Thus, it is applied in every layer so the input will be related to a relevance map $R_p^{(1)}$ in which the value of $\sum_p R_p^{(1)}$ is equal to the relevance value in the output $f(x)$, satisfying the conservation property.

The LRP formulation extends to two other variants:

- ϵ-variant

  Although the standard method conserves the relevance in the propagation process, the explanation can become sensitive to noise. This is because the denominator in the expression can become very small due to canceling out of positive and negative values, resulting in high values on the scores. Therefore, the rule is altered to accept an $\epsilon > 0$ so it gains better numerical properties, as defined bellow.

  $$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j} + \epsilon \; sign(\sum_{i'} z_{i'j})} R_j^{(l+1)}$$

- β-variant

  This variant sets a β value which regulates the impact of negative contributions when backwarded in the relevance. These negative values are considered inhibitors and its increase results in relevance maps restricted to only the most important regions. The equation, then, is given as

  $$R_i^{(l)} = \sum_j \left(\alpha \times \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+} + \beta \times \frac{z_{ij}^-}{\sum_{i'} z_{i'j}^-} \right) R_j^{(l+1)}$$

  where z_{ij}^+ are the positive and z_{ij}^- the negative contributions of z_{ij}, such that $z_{ij}^+ + z_{ij}^- = z_{ij}$. It requires that $\alpha + \beta = 1$, $\alpha > 0$, $\beta \leq 0$ to ensure that the equation keeps it property of conservation layer-wise.

The applied hyper-parameters behave differently for each trained network. A comparison of the properties of the three approaches is seen in Table 4.1.

	naive ($\epsilon = 0$)	ϵ-variant	β-variant
numerically stable	no	yes	yes
consistent with linear mapping	yes	yes	no
conserves relevance	yes	no	yes

Table 4.1: Comparison of LRP variants

	Average	Median
Gradient	**127.79**	**81.00**
Smooth Grad	44.38	16.00
LRP	66.07	41.00

Table 4.2: Average and Median of the APEM values calculated for each of the explanation methods. The relevance maps based on the Gradient result in the best scores.

4.3 APEM Results

First of all, we analyze the APEM scores for the different methods presented. We used the same 5000 random images in each explanation method and the generated relevance maps, following the visualization techniques presented in Section 4.2. The final hyperparameters used in the Smooth Grad were $n = 100$ and $\sigma = 0.2$, and in the LRP the ϵ-variant were used with $\epsilon = 1$ based on the qualitative analyse of the relevance maps created in experimentation.

Figure 4.1 compares some examples of the final map for the Gradient, Smooth Grad and LRP. Although the relevance images are similar and easily interpretable, there are some particularities in each method. They present higher values in close regions but each one focus in a different part of this region. These differences in focus may result in misunderstandings in a classification explanation.

Then, we calculated and compared their APEM values. In Figure 4.2 we see the boxplots of the APEM for each method and the sets. Also, Table 4.2 presents a summary of the average and median of the applied metric.

Even though the Smooth Grad is directly related to the gradient, it has lower APEM results than the LRP. This might happen because the approach used in the Smooth Grad, which creates less noisy images than the gradient itself, does not consider the input image as a single input but a set of perturbed images with different pixel values. On the other hand, the LRP calculates the relevances to one image and its respective output by considering the parameters and the contributions forwarded in each layer.

The values of ϵ^- and ϵ^+ are also compared separately in the boxplots of Figure

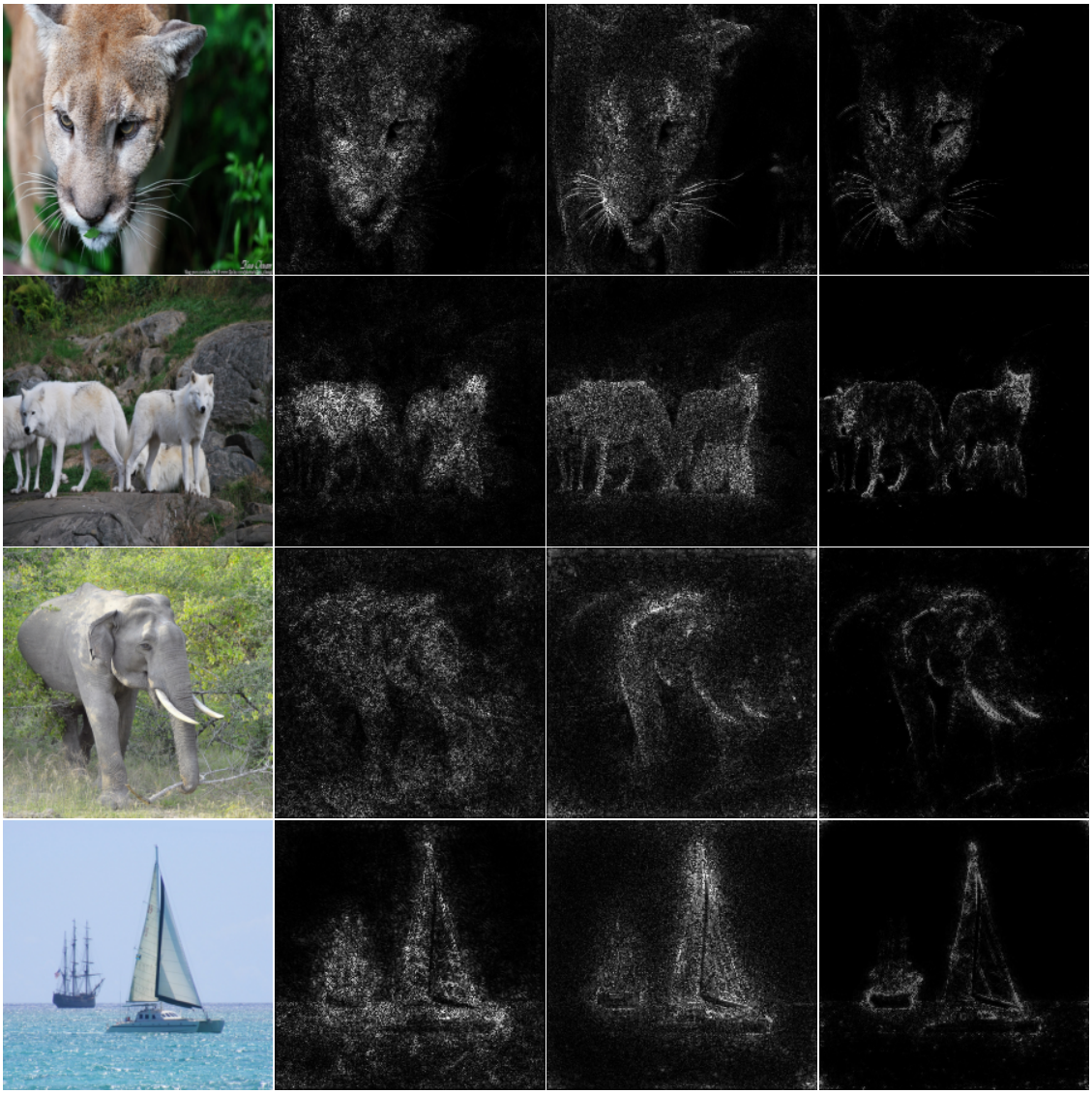


Figure 4.1: Examples of the visualization of relevance maps from the explanation methods. Each row is a different example and each column represents a visualization: original image and the maps calculated from the Gradient, SmoothGrad and LRP (from the left to the right).

4.3. This shows that the methods with the best APEM also have both the minimum and maximum values of ϵ since a correct relevance means a correct irrelevance as well.

We extended the results to a pairwise comparison in Figure 4.4, in which we compared each instance and analyzed the number of times one explanation method beats the other. Once again, it shows that the technique which has a higher total APEM will also be higher for each specific image.

We notice a rare situation in which the Gradient has a lower score than the LRP. Then, we analyzed these examples along with the cases in which the Smooth Grad

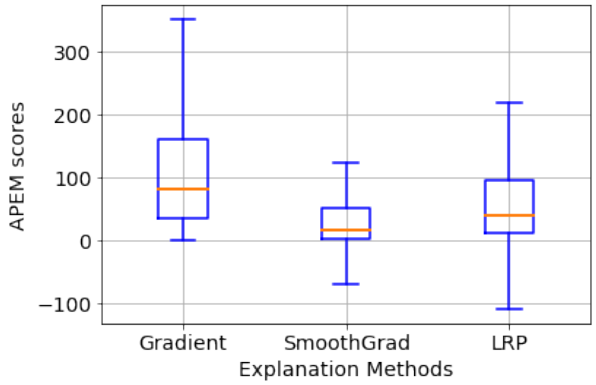


Figure 4.2: Boxplot of the APEM values for each explanation method in the same set of images. Higher values mean better results. The left one is calculated from the Gradient visualization, the SmoothGrad in the middle, and LRP on the right.

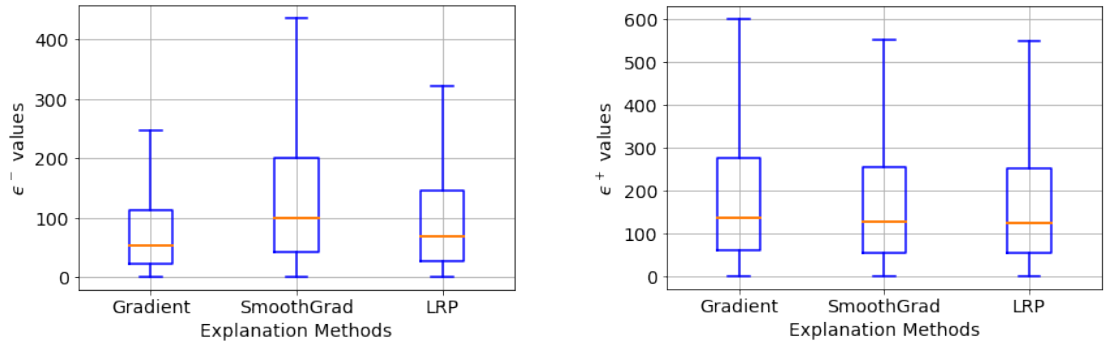


Figure 4.3: Boxplot of the ϵ^- (left) and ϵ^+ (right) values for each explanation method. ϵ^- are expected to have lower values and ϵ^+ higher ones. It shows the results for the Gradient, SmoothGrad, and LRP from the left to the right.

beats the LRP. Table 4.3 shows the values calculated to the APEM of some of the examples in which the dominant method has lower scores. To further analyze these results, Figure 4.5 presents the images relative to the ones in the Table 4.3.

First, by looking specifically at Table 4.3a, the uncommon cases with higher LRP metric scores than with the gradient happens in similar situations. The APEM values calculated to them are slightly different, reaching a maximum distance of 1 in the examples, and the real change is present only in the ϵ^-, keeping the ϵ^+ equal. Moreover, the APEM for these cases are very low and, if compared to their respective images and relevance maps in Figure 4.5, some patterns in the shapes are evident: (*i*) they are all explained by a concise small object centered in the input image, and (*ii*) the map calculated from the gradient is just a little noisier in the region around this object. As both are showing the same regions and the gradient contains just a little more noise, the normalization of the relevance map in the APEM formulation

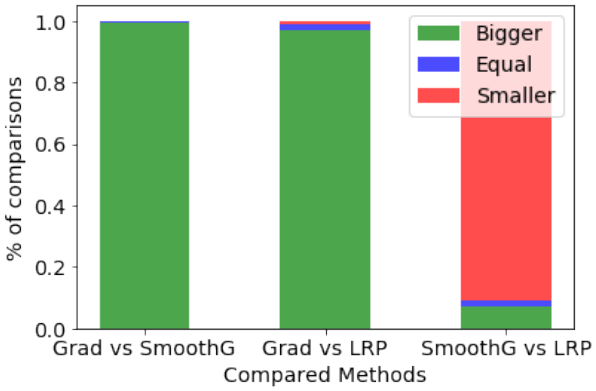


Figure 4.4: Pairwise comparison between explanation methods. It shows the fraction of the total number of compared images in which one technique has a better, equal or worse APEM result than the others for each image in the set.

	Gradient			LRP		
	ϵ^-	ϵ^+	APEM	ϵ^-	ϵ^+	APEM
Image 1	2	4	2	**1**	4	**3**
Image 2	2	3	1	**1**	3	**2**
Image 3	2	6	4	**1**	6	**5**

(a) Gradient vs LRP

	Smooth Grad			LRP		
	ϵ^-	ϵ^+	APEM	ϵ^-	ϵ^+	APEM
Image 4	**275**	**444**	**169**	297	433	136
Image 5	**392**	**470**	**78**	455	467	12
Image 6	**458**	**515**	**57**	482	510	28

(b) Smooth Grad vs LRP

Table 4.3: Comparison of the APEM, ϵ^- and ϵ^+ values of some examples in which the dominant method loses. Recall that the Gradient is the greatest one, followed by the LRP and, lastly, by the SmoothGrad.

makes the pixels perturbation in the correct areas chosen by the LRP a little bit more abrupt. The LRP, then, has a minimal ϵ^- difference with no noticeable changes in the irrelevance, consequently in the ϵ^+. Finally, this oscillation makes the LRP beats the Gradient map in these cases.

Secondly, Table 4.3b shows no apparent pattern and the Smooth Grad beats both in ϵ^- and ϵ^+ in different ways for the examples, which indicates that the relevance maps are actually better for these situations. Comparing these same examples in Figure 4.5,

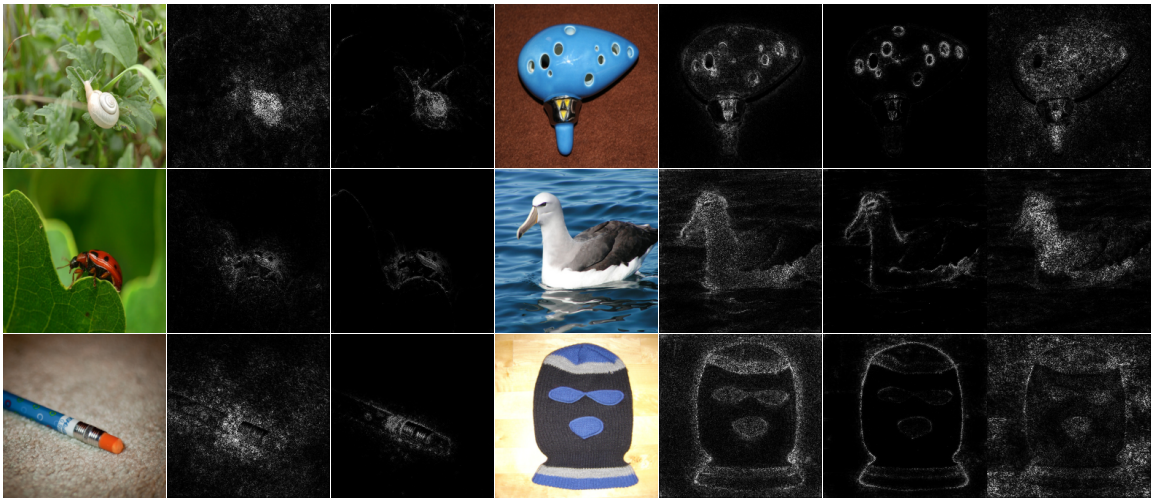


Figure 4.5: Examples of unexpected results. The group of images in the left presents the instances in which the LRP beats the Gradient and each column from the left to the right represents the original images, the relevance maps from the Gradient and the ones from the LRP. On the group on the right, the cases in which the SmoothGrad beats the LRP and the columns continue in the same order with the original image, the maps from the SmoothGrad, LRP and, lastly, the Gradient to serve as a basis for comparison.

we notice that both methods found similar regions in their relevance. However, while smooth grad considered the whole regions as important to the prediction, LRP focused more on the contours and borders of these regions. When compared to the gradient map, it also points to the same regions, agreeing with the other ones, but evidencing the inside of the objects. This supports that, for the cases in which the Smooth Grad beats the LRP, the relevance maps are truly better. This can be a consequence of the model prediction for such cases in which the noise is actually important to the correct output and the LRP might remove some of it.

4.4 Interpretable Visualization vs Explanation

In this section, we extend the debate of the trade-off of human interpretability in visualizations and the actual explainability of a trained model. We address this problem by presenting the relevance raw values of a correct prediction and steps that simplify it until the shapes present in it become more comprehensible. First, the relevance values are summed in the channels dimension so it becomes a relevance map and, then, multiplied by the image so it fits better the shapes in the original image (see [Smilkov et al., 2017]). Figure 4.6 shows some examples of the relevances in each stage of the visualization technique for each method presented here. It is clear that each method

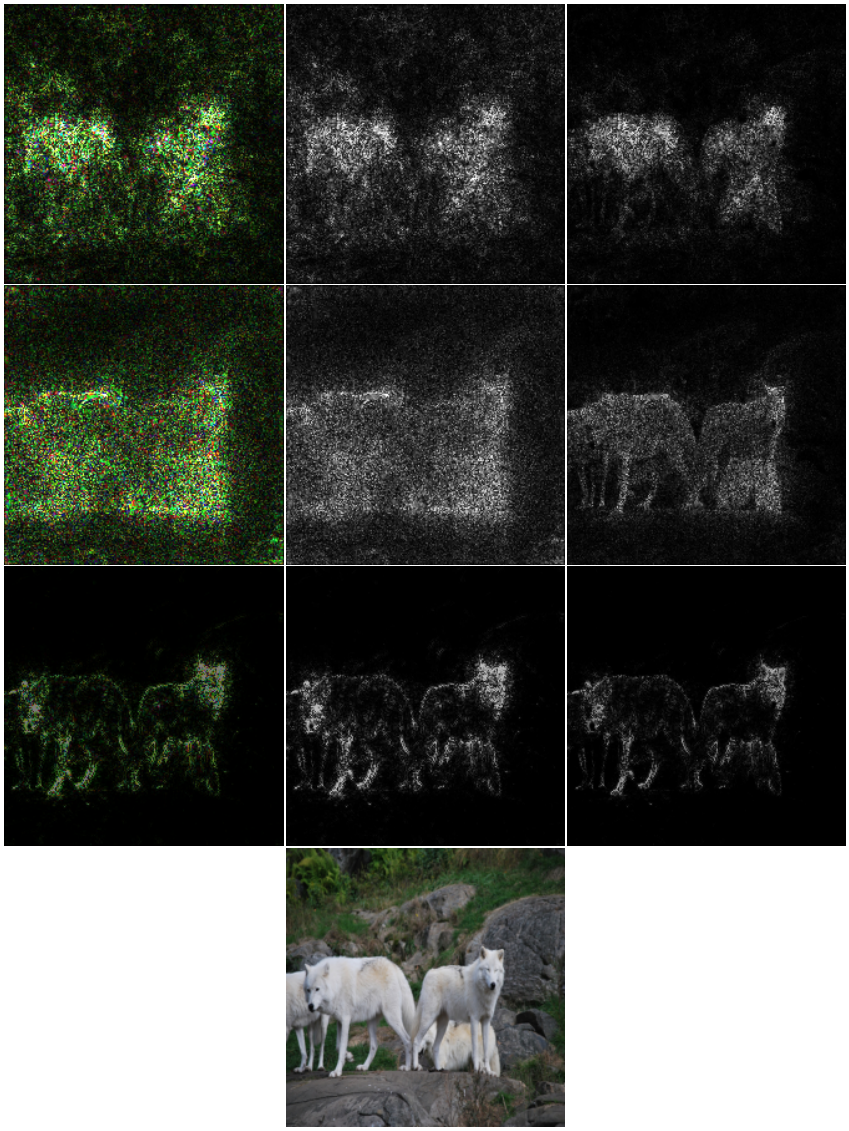


Figure 4.6: Example of the relevance maps for each step in a visualization technique. Each line shows the method being used, which are the Gradient, SmoothGrad and LRP from top to bottom. The visualization steps are presented in each column and can be divided into: (left) relevance in three channels, (middle) mapping into a single channel, and (right) the map multiplied by the original image.

is benefited differently from the processing. Some present great changes in each step, and others, minimal changes.

The applied simplifications are known to lose information, so it is evident that a metric of the explanations after such a process would decrease. Thus, we measure the amount of information that is lost by using the APEM average and median as a quantitative result, shown in Table 4.4. The results follow what the example in Figure 4.6 presents: methods that drastically change the relevance image in the processing

	Average			Median		
	1	2	3	1	2	3
Gradient	**231.76**	176.86	127.79	**137.00**	105.00	81.00
Smooth Grad	**93.40**	74.61	44.38	**41.00**	28.00	16.00
LRP	**101.65**	93.93	66.07	**59.0**	51.00	41.00

Table 4.4: APEM scores of the relevance maps created in each step of the visualization technique for each explanation method. It evaluates the first 3-channel relevance image (1), the mapping to one channel (2) and the resulting image of the multiplication of this final map with the original input image (3).

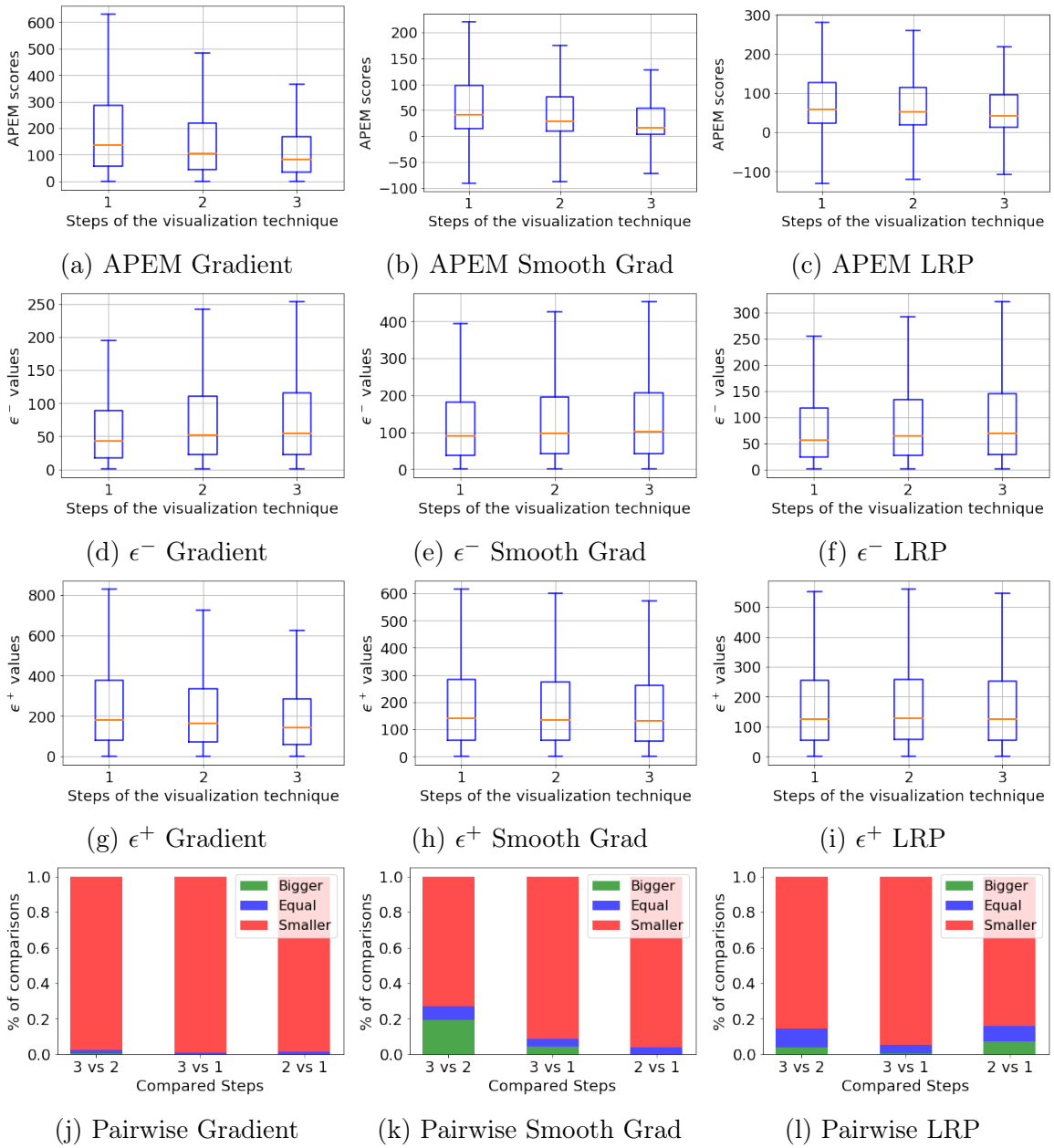
also presents the greatest loss of APEM in general while the ones that only produce minimal change, the lowest loss of metric.

To further investigate the behavior of the APEM in such cases, Figure 4.7 shows the boxplots of the APEM (4.7a, 4.7b, 4.7c), ϵ^- (4.7d, 4.7e, 4.7f), ϵ^+ (4.7g, 4.7h, 4.7i) and pairwise (4.7j, 4.7k, 4.7l) comparisons of each method. It is noticeable the same pattern of improvement: as the APEM increases for a better relevance map, the ϵ^- decreases and the ϵ^+ increases. Moreover, the greatest APEM beats all others in a pairwise comparison. The gradient represents it clearly, in which each simplification of the relevance map makes it have a lower score for every image. On the other hand, even though the same idea happens to all other methods, they present uncommon cases in which a simplification stage can actually enhance their results. This is due to the fact that Smooth Grad and LRP have cases with low APEM score, and a simplification can remove some causes of this decline in value, making it slightly better.

Therefore, the application of the simplifications should consider the decrease in APEM scores compared to the amount of visual comprehension that it brings. Completely noisy images are not as useful as centered clouds of interpretable information, even if it means a loss in explanation. All these factors are important when applying an explanation method in an application.

4.5 Filtering Explanation

Since we presented the behavior of the APEM for various situations, this section expands the use of the metric to filter relevance images to become more interpretable without any loss of essential information. It calculates the maps following the algorithm presented in Section 3.2. First of all, we assume that the gradient, even after transformed into a relevance map, brings a robust map. Their higher values are indeed



(a) APEM Gradient (b) APEM Smooth Grad (c) APEM LRP

(d) ϵ^- Gradient (e) ϵ^- Smooth Grad (f) ϵ^- LRP

(g) ϵ^+ Gradient (h) ϵ^+ Smooth Grad (i) ϵ^+ LRP

(j) Pairwise Gradient (k) Pairwise Smooth Grad (l) Pairwise LRP

Figure 4.7: Results for each step of the visualization technique and explanation methods. It shows the final APEM, ϵ^-, ϵ^+, and pairwise comparisons for the same set of images. (1) is the first 3-channel relevance image, (2) the mapping to one channel and (3) the multiplication of the map with the original input image.

responsible for the classification and the noise created when mapping them into one channel does not make new relevances, even if the APEM reduces. Thus, we first apply the filter technique in the mapping created directed from the gradient after clamped in the 99^{th} percentile. This step is noisier and less interpretable but the APEM is great enough to be considered as a truthful explanation.

Figure 4.8: Examples of the final images from the Filtering Algorithm. It presents the sampled images (top), the relevance maps based on the Gradient before (middle) and after (bottom) the filtering process. Images processed by the algorithm are easier to interpret by an observer.

As we remove noise and the APEM does not drop, the image ends up more understandable and still reliable. Figure 4.8 show some examples of the maps calculated for the images and the last moment before there is a drop in the APEM value. In the example, the filter erased from the relevance map regions outside the main object. This means that the small values attributed to the context, which are often important to the prediction, were actually noise and are not necessary for the visualization. For instance, the grass, stones and water present in the examples had relevance values but they could be removed, leaving only the objects inside it.

The filtering technique can also be used in every other explanation method and in different stages of the visualization process, as shown in Figure 4.9. In the example presented, we used different mappings for the methods, each with its particularities. It resulted in an image which contains only the classified object for every case. Each image focuses on slightly distinct parts of the boat because of the characteristics of the explanation techniques, but all of them removed the relevance related to the sea. Thus,

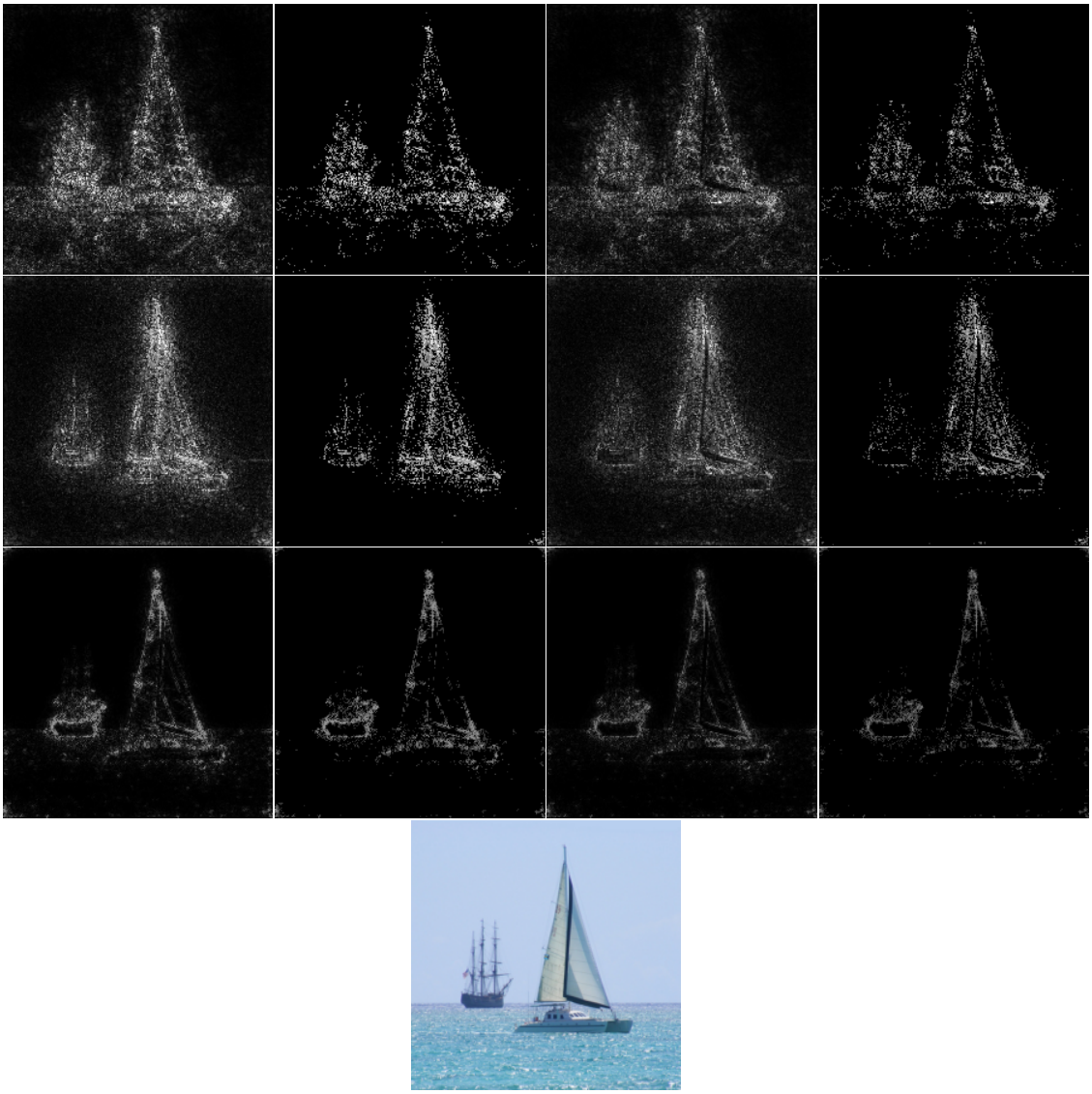


Figure 4.9: Filtering technique applied in each relevance map in the steps of the visualization technique and explanation methods for an example. The rows are the Gradient, SmoothGrad and LRP methods and the columns are, from left to right, the relevance maps and its filtered image, the map multiplied by the original image and its filtered image.

they do not consider the context for this specific classification. Finally, the outcome of the filtering algorithm is easier to see and interpret than the original noisy images.

Chapter 5

Further Analysis

This chapter aims to show some properties of the APEM technique that are important to the reliability of the proposed solution, such as its behavior when using random relevance maps. The properties can also be extended in future works to create new applications in which it is suitable.

5.1 Importance of Irrelevance

One of the contributions of this work is to bring the inverse of the relevance to the metric. This approach prevents the explanation methods from pointing out a few relevant pixels, not giving importance to others which are also relevant. Then, we introduce recall to the metric which already considers precision.

Given an input image, consider the following situations and the behavior of the metric towards them:

- The relevance map calculated results in an image with one small region as important while both this region and another one are actually important: ϵ^- has a low value but the ϵ^+ drastically reduces since there are relevant pixels been considered in its calculation, making the model misclassify the perturbed image faster. Therefore, the metric is penalized by not considering the other relevant region.

- The relevance map calculated results in an image with two regions as important while just one of them is actually important: even though ϵ^+ is not penalized because there is no relevant pixel in the irrelevance map, ϵ^- increases for considering irrelevant regions. The metric is also lower than in a correct relevance map.

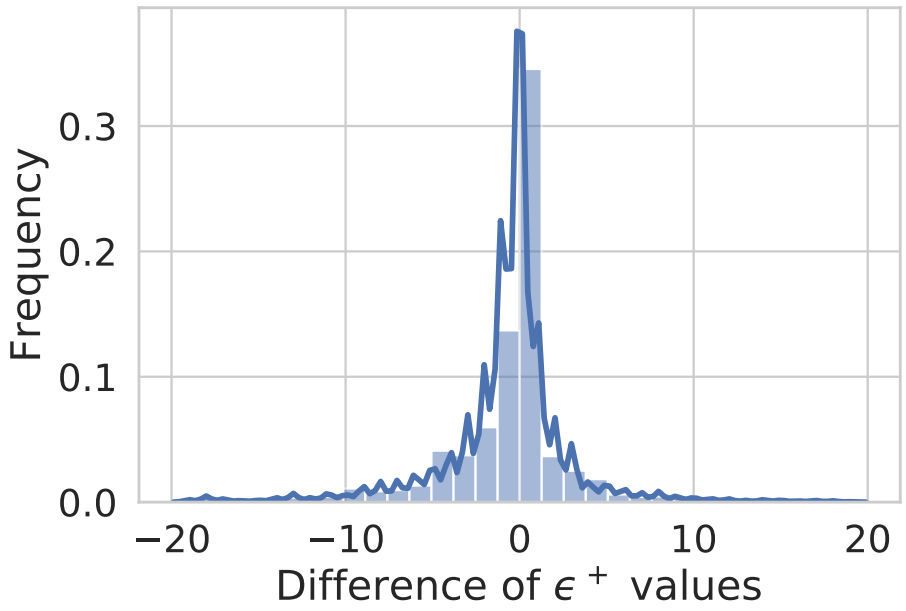


Figure 5.1: Histogram and kernel density estimate (KDE) of the difference between the ϵ^+ of the LRP and the SmoothGrad. It only considers images which have better ϵ^- for the LRP. It has positive values when the irrelevance of the LRP is better and negative otherwise. This means that there are considerable cases for both.

To present this property in our experiments, we compared the results of the LRP with the SmoothGrad. Although LRP beats SmoothGrad in the metric results, we observed that the generated relevance images from the LRP are often more simplified. This can result in lower ϵ^+ values even when the metric is better. Figure 5.1 shows the histogram and the kernel density estimate (KDE) of the difference between the ϵ^+ of the LRP and the SmoothGrad for each image in which the ϵ^- of the LRP has lower values. We note that the irrelevance of the LRP has both greater and lower values than the other one because it also contains relevant pixels in some instances, even if the metric itself is better and their ϵ^- distributions resemble.

5.2 Random Relevance Maps

In order to test the robustness of the metric, we use the same 5000 images and their respective relevance maps evaluated in Section 4.3. However, in this experiment, we shuffle the relevance and irrelevance images to create random images based on the same distribution given by the methods. Random relevance maps are supposed to have the same ϵ^- and ϵ^+, as the relevance values are now high in both true relevant pixel and in non-relevant. For each image, we created ten shuffled maps and if the

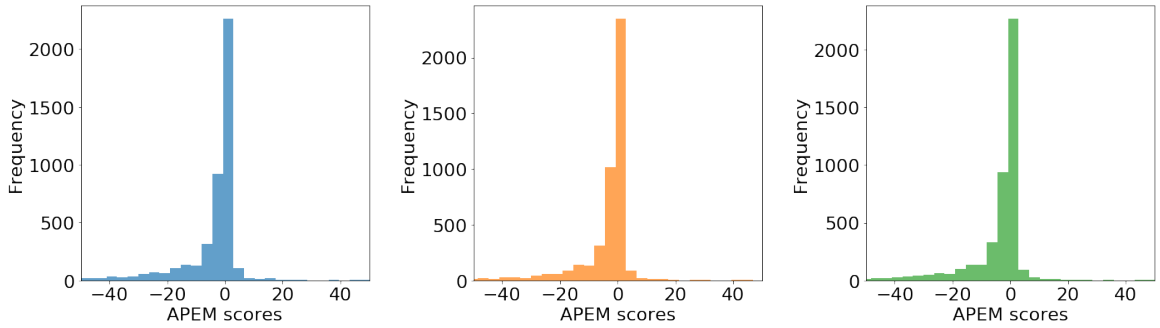


Figure 5.2: Histogram of APEM values for shuffled relevance maps of different stages of the visualization technique based on the Gradient. It oscillates around zero, meaning that bad explanation methods will also have results close to zero.

model misclassifies one of them, it sets the ϵ^- and ϵ^+. First of all, we tested this for the different stages of the visualization of the Gradient method and Figure 5.2 shows a histogram of the APEM calculated for them.

Then, we evaluate the average and median using the same process for every method. The average oscillates around 0, resulting in small values, and the median is 0 for every case tested. Since the shuffle is presented as reliable, it can be used to normalize the original APEM values of the images. Therefore, we can calculate the normalized APEM_{norm} for the shuffled $\epsilon\text{-}shuffle^+$ and $\epsilon\text{-}shuffle^-$ values as

$$APEM_{norm} = \frac{\sum_{i=1}^{n} \left(\frac{\epsilon_i^+}{\epsilon\text{-}shuffle_i^+} - \frac{\epsilon_i^-}{\epsilon\text{-}shuffle_i^-} \right)}{n}$$

Recalculating the results presented in Section 4.3 with the new normalized APEM_{norm} results in the values in Table 5.1. As $\epsilon\text{-}shuffle^+$ and $\epsilon\text{-}shuffle^-$ are known to be equal or very close, and their values are between the maximum and minimum ϵ, represented by ϵ^- and ϵ^+, the normalization keeps the same properties of the original APEM but in a space where the values are useful for visualization and comparisons.

5.3 Misclassified Images

Until now, there was only correctly classified examples. However, we have to understand how the metric deals with misclassified images to properly explain and debug a model. In order to analyze the wrong classifications, we evaluated the APEM in the same process for a set of 5000 random misclassified images. In this case, the label we used to create the relevance maps were the ones predicted by the model.

	1	2	3
Gradient	**1.11**	**0.85**	**0.66**
Smooth Grad	0.39	0.30	0.21
LRP	0.53	0.48	0.40

Table 5.1: $APEM_{norm}$ calculated from shuffled values of the relevance maps of each explanation method in the steps of the visualization technique. The steps are the first 3-channel relevance image (1), the mapping to one channel (2) and the resulting image of the multiplication of this final map with the original input image (3). As the values are shuffled in the maps, we expect the scores to be as closer to 0 as possible.

	Average	Average Misclassified	Median	Median Misclassified
Gradient	**127.79**	**43.58**	**81.00**	**25.00**
Smooth Grad	44.38	12.65	16.00	4.00
LRP	66.07	20.34	41.00	11.00

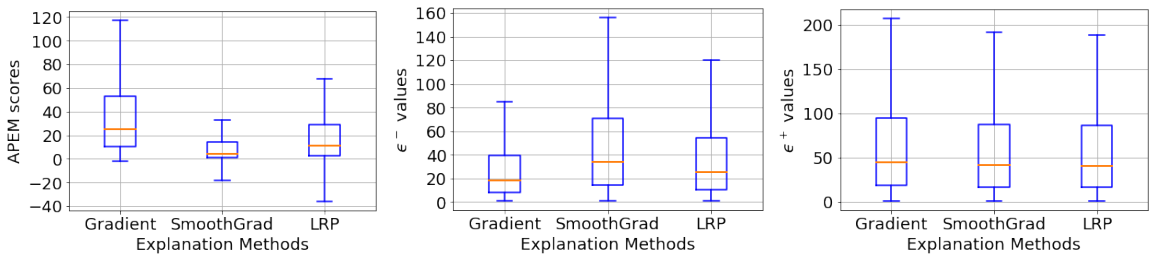
Table 5.2: Average and Median of the APEM values for correctly classified and misclassified images. Maps are calculated considering the prediction as the ground truth. Misclassification leads to an overall reduction in the scores.

As the model is mostly uncertain about a decision when it predicts a wrong label, the model does not need great perturbations to make him change a prediction, resulting in lower ϵ values. This reduction leads to a lower APEM score. On the other hand, this should not influence the quality of the explanation, disregarding the understanding of the generated relevance map. To put simply, the overall scores reduce but the best methods have to keep their ranking positions in the comparisons since they are still explaining an output for a given model and an input instance.

Table 5.2 shows the average and median for their APEM values and compares with the ones of the correctly classified images. The reduction in the score is evident and the order is kept unchanged. Figure 5.3 presents the boxplots which support these results, indicating that their distributions are similar.

5.4 Correlation of APEM and Loss

Another property investigated is the correlation between the APEM and the loss in the prediction. For this, we used the same datasets of correctly classified images, misclassified images, and the total set which contain both of them. Recall that the relevance map calculated for the misclassified images is based on the model prediction even though the loss uses the ground truth. We also compare the greatest probability



(a) APEM (b) ϵ^- (c) ϵ^+

Figure 5.3: Boxplot of the values of APEM, ϵ^- and ϵ^+ for misclassified images. The metric is calculated as if the first prediction was the ground truth and the Gradient, SmoothGrad and LRP methods are applied. This presents the behavior of the metric toward images in which the model is not able to classify correctly.

for the class in the prediction to the Loss, which we refer to as confidence.

The correlation method used was the Spearman's rank correlation coefficient [Croux and Dehon, 2010] because of the non-linear relationships present in the data. This correlation is equal to the Pearson correlation between the rank values of the variables. A correlation close to +1 occurs when the observations have a similar rank between the variables and to -1 when they have a dissimilar one.

Table 5.3 shows the correlations and the statistical significance of them. We note some characteristics present in the results:

- Correctly classified images have higher correlations while the misclassified images have virtually none.
- Observations have a dissimilar rank between the APEM and Loss, resulting in a negative correlation.
- The methods which give greater APEM values also are the ones with higher correlation.

To conclude, the best explanation methods in terms of APEM score have a higher correlation with the Loss. Furthermore, great APEM values mean lower losses. Thus, good explanations along with the APEM calculation can be used to tell if the output of a model is reliable.

	Loss (Correctly Classified)	Loss (Misclassified)	Loss (Total)
Gradient	-0.827†	0.069†	-0.543†
Smooth Grad	-0.470†	-0.025	-0.349†
LRP	-0.602†	0.001	-0.446†
Confidence	-1.000†	-0.121†	-0.697†

Table 5.3: Spearman correlation between the APEM score for each explanation method and the loss of the evaluated model ($\dagger$ represents statistical significance with $\rho < 0.01$). The correlation of the confidence of the most probable class and the loss is also presented for comparison. Both the correctly classified and misclassified images are evaluated along with the union of them.

Chapter 6

Conclusions

In this work, we proposed the Adversarial Perturbation Explanation Metric (APEM), a robust metric which evaluates the quality and reliability of explanation methods. This approach compares such methods quantitatively and qualitatively, avoiding visual inspection. Moreover, it considers every relevance value for an input image to create perturbations and the irrelevance map to guarantee that no relevant pixel is left out.

We presented a comparison of some well-known explanation methods using our metric. Along with it, some characteristics of the methods and how the metric react to it. Furthermore, we showed some properties which make the metric robust to the results. First, we showed the importance of using irrelevance as the result varies if the relevance map is completely precise but it is omitting other relevant pixels. Then, we analyzed the responses to random relevance maps and misclassified images, showing that the metric presents a zero value in cases of randomness and the metric drastically falls when the model is not able to correctly predict an instance. Finally, we correlated the metric results to the output of the model in different situations.

It was also proposed in this work an study on the techniques of visualization and the effect of its simplification into the reliability of the resulting images, and a technique which works around this problem. The technique is one of the applications in which APEM values serve as a tool. It filters the relevance maps into more interpretable maps while keeping all the essential information. The proposed algorithm can be used after every explanation method to create less noisy images and facilitate its understanding by an observer.

A measure of reliability of the explanation methods allows users to choose the appropriate option for a task. We mainly present a metric constrained to the data set and the trained model of an application and some primary properties. Many other important properties are worth exploring so we can understand more about the metric. After

that, we can investigate the possibility for a global range of the measurement independent of the evaluated data set. This brings the possibility of calculating thresholds of values which assert the overall quality of a metric without the need for comparisons. The aim of this work is to introduce a way to help in the development and investigation of the explanation methods.

Bibliography

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 9525--9536.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803--1831.

Binder, A., Bach, S., Montavon, G., Müller, K.-R., and Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. In Kim, K. J. and Joukov, N., editors, *Information Science and Applications (ICISA) 2016*, pages 913--922, Singapore. Springer Singapore.

Cadamuro, G., Gilad-Bachrach, R., and Zhu, X. (2016). Debugging machine learning models. In *ICML Workshop on Reliable Machine Learning in the Wild*.

Croux, C. and Dehon, C. (2010). Influence functions of the spearman and kendall correlation measures. *Statistical Methods and Applications*, 19(4):497--515.

Dosovitskiy, A. and Brox, T. (2016). Inverting visual representations with convolutional networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4829--4837. IEEE Computer Society.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*.

Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3449--3457. IEEE Computer Society.

Freitas, A. A. (2013). Comprehensible classification models: a position paper. *SIGKDD Explorations*, 15(1):1--10.

Girshick, R. B. (2015). Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440--1448. IEEE Computer Society.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580--587. IEEE Computer Society.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

Goodman, B. and Flaxman, S. R. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50--57.

Hartford, J. S., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In Precup and Teh [2017], pages 1414--1423.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *CoRR*, abs/1707.01154.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K., and Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2912--2920. IEEE Computer Society.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8595--8598. IEEE.

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., and Feng, D. (2014). Early diagnosis of alzheimer's disease with deep learning. In *IEEE 11th International Symposium*

on Biomedical Imaging, ISBI 2014, April 29 - May 2, 2014, Beijing, Chin, Beijing, China, pages 1015--1018. IEEE.

Mahendran, A. and Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 5188--5196. IEEE Computer Society.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236--1246.

Montavon, G., Braun, M. L., and Müller, K. (2011). Kernel analysis of deep networks. *Journal of Machine Learning Research*, 12:2563--2581.

Montavon, G., Samek, W., and Müller, K. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1--15.

Nguyen, A. M., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 427--436. IEEE Computer Society.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

Payan, A. and Montana, G. (2015). Predicting alzheimer's disease - A neuroimaging study with 3d convolutional neural networks. In De Marsico, M., Figueiredo, M. A. T., and Fred, A. L. N., editors, *ICPRAM 2015 - Proceedings of the International Conference on Pattern Recognition Applications and Methods, Volume 2, Lisbon, Portugal, 10-12 January, 2015.*, pages 355--362. SciTePress.

Precup, D. and Teh, Y. W., editors (2017). *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*. PMLR.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91--99.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram, B., Shah, M., Smola, A. J., Aggarwal, C. C., Shen, D., and Rastogi, R., editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135--1144. ACM.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211--252.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learning Syst.*, 28(11):2660--2673.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. (2016). Grad-cam: Why did you say that? *CoRR*, abs/1611.07450.

Shen, D. and Suk, G. W. H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248. PMID: 28301734.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. A. (2014). Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In Precup and Teh [2017], pages 3319--3328.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2013). Intriguing properties of neural networks. *CoRR*, abs/1312.6199.

Yosinski, J., Clune, J., Nguyen, A. M., Fuchs, T. J., and Lipson, H. (2015). Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818--833. Springer.