Assessing the Reliability of Visual Explanations of Deep Models with Adversarial Perturbations

Dan Valle
CS Dept@UFMG & Wildlife Studios
Belo Horizonte, Brazil
dan.valle@wildlifestudios.com

Tiago Pimentel
CS Dept@UFMG & Kunumi
Belo Horizonte, Brazil
tpimentelms@gmail.com

Adriano Veloso
CS Dept@UFMG
Belo Horizonte, Brazil
adrianov@dcc.ufmg.br

Abstract—The interest in complex deep neural networks for computer vision applications is increasing. This leads to the need for improving the interpretable capabilities of these models. Recent explanation methods present visualizations of the relevance of pixels from input images, thus enabling the direct interpretation of properties of the input that lead to a specific output. These methods produce maps of pixel importance, which are commonly evaluated by visual inspection. This means that the effectiveness of an explanation method is assessed based on human expectation instead of actual feature importance. Thus, in this work we propose an objective measure to evaluate the reliability of explanations of deep models. Specifically, our approach is based on changes in the network's outcome resulting from the perturbation of input images in an adversarial way. We present a comparison between widely-known explanation methods using our proposed approach. Finally, we also propose a straightforward application of our approach to clean relevance maps, creating more interpretable maps without any loss in essential explanation (as per our proposed measure).

Index Terms—Deep Networks, Explainability, Interpretability

I. INTRODUCTION

The development of deep neural networks for computer vision applications is at the crossroad of two major trends. The first trend is associated with increasingly complex models leading to state-of-the-art performance in applications that vary from healthcare [1], [2] to economics [3], to ranking [4]. The second trend is the increasingly perceived importance of transparency and accountability in a range of applications.

Deep neural networks do not provide insights into their complex behavior, and thus several methods attempt to unveil the factors contributing most to these networks black-box decisions [5]–[8]. These explanations are important to identify potential bias/problems in the training data [9], to ensure compliance to existing regulations, and guarantee the model performs as expected [10]. This leads to an increase in interpretability [11], making models more transparent and their predictions more understandable.

The explanation of deep models for computer vision applications is usually given in terms of interpretable visualizations of the relevance of pixels from input images (i.e., pixel

relevance maps). Currently, the reliability of these explanations is largely assessed by visual inspection, and thus it is likely to be the case that existing explanation methods are being evaluated based purely on human expectation rather than on actual feature importance [12]. As deep neural networks have special behaviors [13] and can be easily confused [14]–[16], intuitive visualizations can be misleading and different from the real importance given to features in a trained network.

An objective measure from which the quality of model explanations can be systematically assessed is largely lacking. In this paper we propose an approach for comparing the reliability of explanations. Our approach calculates a reliability measure based on changes in the model outcome resulting from adversarial perturbations of input images. Our specific contributions are summarized as follows:

- We introduce Adversarial Perturbation Explanation Measure (APEM), which evaluates pixel relevance maps by assuming model decisions must depend on its explanation [17], [18]. Therefore, the most relevant features in the input image are the ones that influence the most the output of the model. To calculate APEM, we probe the model outcome by perturbing the input image based on a relevance map, multiplied by the sign of the gradient, trying to maximize the error with minimal input modification.
- We compare a variety of *state-of-the-art* methods used for assessing feature importance in a controlled and standardized setting. We show how common practices in these explanation methods abdicate important information impacting models decisions in order to make visualizations more comprehensible and visually meaningful.
- We show that it is also important to consider the low scores of relevance in order to avoid privileging explanation methods whose maps are restricted to a few concentrated high values. Thus, we use *irrelevance* maps together with relevance ones, and we show that this balances the 'precision' and 'recall' of the resulting maps.
- We present a new approach based on the proposed measure to clean relevance maps, making them more intuitive and understandable while keeping explanations reliable. This results in images with minimum noise while ensuring that no important explanation is lost.

II. Background and Related Work

In this section we give an overview of methods focused on explaining deep network outcomes. Then, we present studies reporting how visualizations can be unreliable in some cases. Finally, we also discuss approaches for evaluating the quality of relevance maps generated by existing explanation methods.

A. Explaining Deep Neural Network Decisions

The large number of layers employed by deep models combined with their non-linearities makes it difficult to identify what is being considered in each decision. Some works attempt to understand individual neurons in deep neural networks by creating visualizations from higher level features in Autoencoders [19], [20] and Deep Belief Networks [19]. These studies show how models can find patterns that are similar to what humans consider relevant in the domains analyzed.

Understanding what is important to the performance of deep models is essential to find situations in which they fail and to improve them. Authors in [21] created a visualization for each input image, which shows the patterns in the input that resulted in a specific activation on further layers. Other techniques tackle this problem by inverting the input images representations and analyzing their remaining information [22], [23].

Authors in [24] presented a gradient-based approach to interpret decisions of Convolutional Neural Networks (CNNs). They proposed saliency maps that are computed using the gradient of each class' score. This method generates visually noisy relevance maps, though. Smooth Grad [25] builds on their work and tackles the issue with noisy images. In order to do so, the authors created n noisy copies from each input image and average the relevance maps calculated for them. Other works also use the gradient in different ways to achieve contrasting visualization results [26]–[28]. LIME [6], on the other hand, explains predictions by learning an interpretable model locally around the input image.

A distinct group of model explanation techniques calculates relevance scores for pixels by applying a "backpropagation" method. It computes the relevance from the output back to the input image, assuring a layer-wise conservation property [17], [29]. Consequently, neurons that contribute more to the ones in the following layers have higher scores. Authors in [30] extended this method to Fisher Vectors, showing how distinct models may consider certain regions of images as relevant or not. In addition, their conclusions shed doubts on the reliability of model decisions which can be biased by context.

B. Reliability of Visualizations

Deep neural networks are complex systems which are not fully understood yet [14], [16], [31]. Authors in [31] and [16] investigated the counter-intuitive property called Adversarial Examples, which is further explored in this work. They showed that deep models decisions can be fragile, being susceptible to directed perturbations that are imperceptible to humans. The nonlinear nature of neural networks is presented as the main cause of such vulnerability.

Authors in [14] further analysed this subject, questioning the differences between patterns considered by humans and models. They create noisy images which received extremely high confidence model decisions. In this work, we build on their line of questions and ask if state-of-the-art explanation methods capture the actual importance of input features to a specific model, or if they trade this off for patterns that are more intuitive to humans. Authors in [12] addressed this question and presented two tests to evaluate the reliability of explanation methods:

- **Model Parameter Randomization Test:** Compares the explanation maps of a trained model with the ones of a randomly initialized untrained network of the same architecture. If the results of the two cases do not differ significantly, it means that the explanation maps are insensitive to the parameters of the model.
- **Data Randomization Test:** Compares the explanation 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 the labels were randomly permuted. If the results do not differ significantly, the explanation maps are insensitive to the relationship between the input images and the original labels.

They evaluated widely used explanation methods, and concluded that visual inspection is a poor way of evaluating explanation results. These results imply the need for techniques which compare methods in a more objective manner, based on proper estimates of feature importance.

C. Evaluation of Explanation Methods

As stated above, comparisons between explanation methods of deep networks are commonly qualitative. They usually compare examples of relevance maps, preferring the ones more correlated to human expectation. In this sense, we consider the work in [18] as the closest to ours. The authors provided a quantitative measure that evaluates how the relevant areas affect the correct prediction of a given model. While they consider block regions of relevance and apply random changes to it, we use small guided perturbations in the whole input image. We evaluate the magnitude necessary for these perturbations to make the model change its output, considering that better relevance maps should require smaller magnitudes.

Other relevant work is [32], in which the authors discussed the desirable properties of explanations and possible evaluation metrics, which they defined as follows:

- **Explanation Continuity:** Ensures that if two inputs are nearly equivalent, then the explanations of their predictions should also be nearly equivalent.
- **Explanation Selectivity:** More relevant features should have stronger impacts on the classification. Thus, if features are attributed relevance, removing them should reduce evidence of the output.

Finally, authors in [17], [18] created concepts based on the quantification of this second property, measuring how fast an evaluated function starts to decrease when removing features with the highest relevance scores.

III. APEM: COMPARING VISUAL EXPLANATIONS WITH ADVERSARIAL PERTURBATIONS

We tackle the quantitative evaluation of explanations by performing guided perturbations to the original images. Typical explanation methods generate one relevance value for each pixel, resulting in images with the same dimensions of the input image, but in which values represent feature relevance $R = [[r_{i,j} | 0 \leq r_{i,j} \leq 1]]$. Thus, higher $r_{i,j}$ values imply those pixels have higher influence in the model decision.

A. Assessing Reliability by Comparing Relevances

Once a model is trained, the relevance maps can be calculated for a set of input images based on the model parameters. Our aim is to produce scores that can be ranked to compare possible explanations for a given model, giving higher values to features that impact more the model decision.[1]

We assume that changes in pixels associated with higher relevance values should impact the output more. In contrast, changing the irrelevant pixels should result in smaller impacts on the model outcome [18]. These perturbations are hardly noticeable and follow the directions of larger impact to the model decision, i.e. the models' gradients [16]. We first normalize relevance values in R by its l_1-norm: $R_{norm} = R/norm(R)$ so that explanation methods with different relevance scales could be directly compared. Then, we create a directed relevance R_{dir} as show in Equation 1. Specifically, for a model with parameters θ, an input image x and model output $\hat{y} = f_\theta(x)$ associated with x, we use the sign of the gradients from the loss function $J(\theta, x, \hat{y})$ to direct the relevances:

$$R_{dir} = R_{norm} \odot sign(\nabla_x J(\theta, x, \hat{y})) \quad (1)$$

Further, the R_{dir} which correctly represents the model relevance would need a minimum perturbation to maximize the deviation from the model decision, as it is equivalent to one gradient ascent step in the pixel space. Therefore, there is a minimum ϵ^- value which makes the model change its prediction for the perturbed image x' created from x:

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

Similarly, we argue that the values of irrelevance can analogously be extracted from R by making $R^I = 1 - R$. Moreover, the best R_{dir}^I calculated from R^I should take longer to change model decisions, as perturbations in irrelevant pixels should not cause significant changes in the output. This results in an ϵ^+ value and, consequently, a gap between it and ϵ^-. APEM is finally given as the average of all the gaps for a set of n images, as expressed in Equation 2. The entire process of computing APEM is depicted in Figure 1.

$$\text{APEM} = \frac{\sum_{i=1}^{n} (\epsilon_i^+ - \epsilon_i^-)}{n} \quad (2)$$

---

[1]APEM, however, does not produce results in the same scale for different models, so it cannot be straightforwardly used to compare explanations for different models with diverse capacities. Our objective is, thus, to compare different explanation methods given a fixed model.
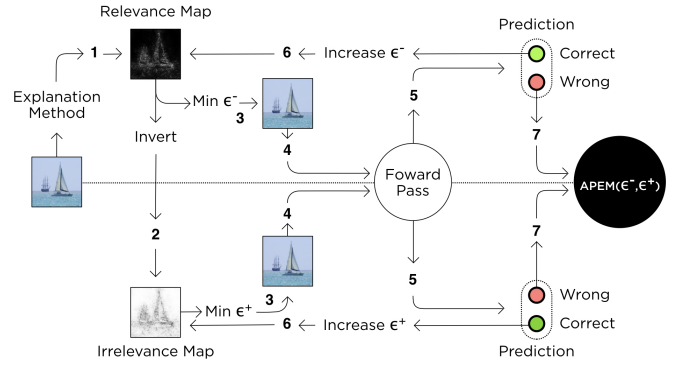

Fig. 1: Calculating APEM for an arbitrary input image. The transitions are enumerated sequentially from the first step to the last, and the ϵ values are stored and modified throughout the process. The diagram shows the relevance and irrelevance maps for an input image, and the gradient was used to create relevance and irrelevance images. The final steps consist of perturbing the original image by using the maps until the model changes its output.

B. An Algorithm to Make Relevance Interpretable

Since the reliability of a relevance map can now be measured, we may use APEM to assess if the quality of explanations was reduced when an explanation method simplifies relevance maps to make them more interpretable. We argue that a good simplification is one which does not reduce the APEM value of the original map. As discussed, pixels which are less relevant to the model decision have smaller values in the corresponding relevance map. When their relevance values are zeroed, their irrelevances are consequently set to 1 and APEM can be calculated to the new map. If the APEM value remains unchanged, the pixels are considered to have no influence in the prediction and can be safely excluded from the explanation. This makes explanations more understandable, while keeping the same reliability levels.

We propose here, thus, a simple algorithm to filter relevance maps, removing the noise in it so humans can more easily interpret them. Basically, it zeroes the least relevant non-zero values in a relevance map iteratively, until any change in the relevance map causes a reduction in the APEM value. This algorithm can be applied to any explanation map calculated for an image and a trained model. While other methods have to deal with the trade-off of losing information to make explanations and visualizations more interpretable, the one presented here works as a way to enhance interpretation while not reducing explainability. Therefore, we have reliable visualizations and explanations which are also easier to understand.

IV. EXPERIMENTS

In this section we report results from the comparison of different explanation methods using APEM. We also show how more interpretable visualizations may affect the actual explanation and how we can effectively tackle this problem,

creating meaningful visualizations while keeping the same APEM values.

A. Model, Data and Explanation Methods

Our model is a VGG-16 [33] trained on the ILSVRC2012 dataset [34] using PyTorch [35]. We used two sets of 5,000 random images each – one of correctly classified and another of misclassified images. These were taken from the validation set and used to create the relevances for each explanation method to be evaluated. We compare six explanation methods:

- **"Pure" gradients**: the gradients are simply interpreted as a relevance map.
- **Smooth Grad** [25]: it smooths the gradient, by applying a Gaussian kernel, instead of the raw gradient. This results in a sensitivity map M. Then, it averages the sensitivity maps using random samples obtained from the neighborhood of an input image x, formulated as:

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

  where n is the number of samples used, and $\mathcal{N}(0, \sigma^2)$ the Gaussian perturbation with standard deviation σ.

- **Layer-wise Relevance Propagation (LRP)** [29]: it computes the relevance of the pixels of an input image by considering their impact on the output of the model. LRP uses a graph structure 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.
- **Guided Backpropagation** [26]: it corresponds to the gradient method in which negative gradient entries are set to zero while backpropagating through a *ReLU* unit.
- **Grad-CAM** [27]: it computes the relevance map as the gradient of the class score with respect to the feature map of the last convolutional unit of the network.
- **Guided Grad-CAM** [27]: Grad-CAM combined with Guided Backpropagation through an element-wise product for pixel-level granularity.

For each of the aforementioned explanation methods, we get relevance values R and clamp them to an upper-bound using its ninety-ninth percentile ($r'_{i,j} = \min(r_{i,j}, R_{99})$). Then, we multiply the new relevance values by their respective original image pixel values ($R'' = R' \odot I$, where $\odot$ is the element-wise product). This results in a cleaner visualization as proposed in [25]. As we are using RGB images, we reduce the number of channels in our relevance maps by summing them and normalizing it to the range $[0, 1]$.[2]

B. APEM Results

We start our analysis by showing a comparison between the different explanation methods using APEM values. Hyper-

---

[2]Though we present main results for all explanation methods described here, we will mostly focus on the Gradient, Smooth Grad and LRP methods throughout this work. We selected these three methods because of their different approaches and outcomes.
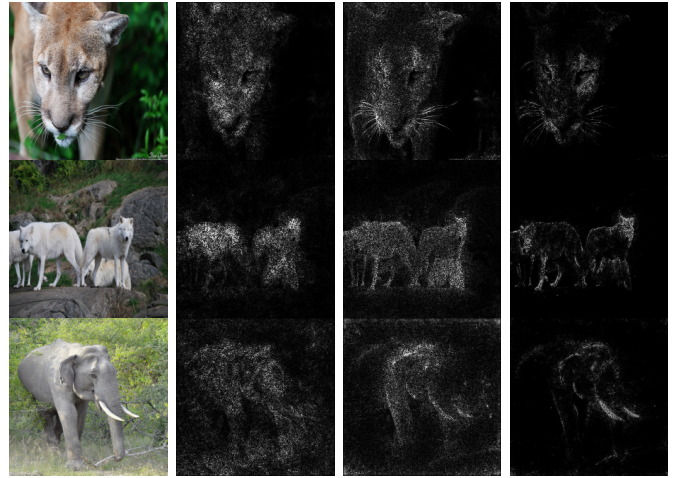

Fig. 2: Examples of visualizations of relevance maps obtained by different explanation methods. Each row corresponds to an input image and each column shows a visualization: original image, and then the maps obtained with Gradient, SmoothGrad and LRP methods, respectively.
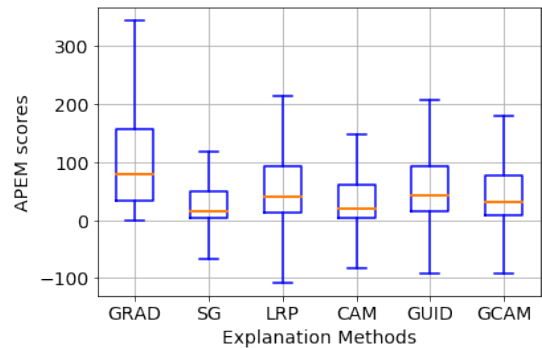

Fig. 3: Boxplots showing APEM values for each explanation method in the same set of images. They represent the Gradient, Smooth Grad, LRP, Grad-CAM, Guided Backpropagation and Guided Grad-CAM from left to right.

parameters used in Smooth Grad were $n = 100$ and $\sigma = 0.2$, and for LRP we used its ϵ-variant with $\epsilon = 1$.

Examples of the final maps are shown in Figure 2, for the Gradient, Smooth Grad and LRP methods. Although the relevance maps seem similar and easily interpretable, there are some particularities that are worth mentioning. In particular, while all visualizations present higher values in close regions, each visualization focuses in different parts of a same region. Figure 3 shows APEM boxplots for each explanation method when considering all 5,000 images. Higher APEM values mean better results, and clearly, the distribution of APEM values differs greatly depending on the explanation method. This indicates that while visualizations seem similar, they might not express the actual feature importance.

In order to assess the stability of the APEM values for each method across different network architectures, Table I presents

TABLE I: Median of the APEM values calculated for each explanation method for VGG and ResNet models. The relevance maps based on the Gradient result in the best scores. Implementing LRP on ResNet is not trivial, so we did not calculate the APEM score for this model.

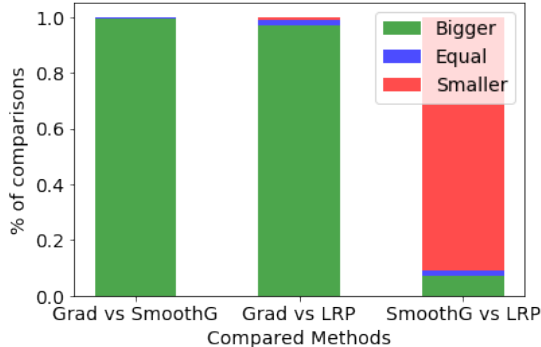|	VGG	ResNet
Gradient	**81.00**	**62.00**
Smooth Grad	16.00	15.00
LRP	41.00	–
Guided Bp	43.00	28.00
Grad-CAM	20.00	21.00
Guided Grad-CAM	31.00	19.00

Fig. 4: Pairwise comparison between explanation methods. It shows the fraction of the total number of images in which one explanation method has a better/equal/worse APEM performance.

the median of the results seen in the boxplots compared to a ResNet [36] evaluated in the same conditions. We noticed the ranking of most explanation methods maintains the same order. The only exception is Grad-CAM and its guided variant, which show very similar results when evaluated on ResNet.

These results can be extended to a pairwise comparison, as shown in Figure 4. We counted the number of input images for which one explanation method beats the other in terms of APEM performance. Both Figures 3 and 4 show LRP achieves a better APEM performance than SmoothGrad, but it also presents a higher standard deviation. This indicates that LRP is usually better than SmoothGrad, but it presents worse results for a few input images.

C. Interpretable Visualizations vs. Actual Explanations

In this section we extend our discussion about the trade-off between a relevance map being interpretable and the actual importance that features within the relevance map have to model decisions. We address this problem by presenting the raw relevance values of a correct prediction and standard filtering steps that simplify the visualization until it becomes more comprehensible. First, relevance values are summed in the channel dimension so it becomes a relevance map. Then, they are multiplied by the grayscale image, so that it better fits the shapes in the original image, as proposed in [25].
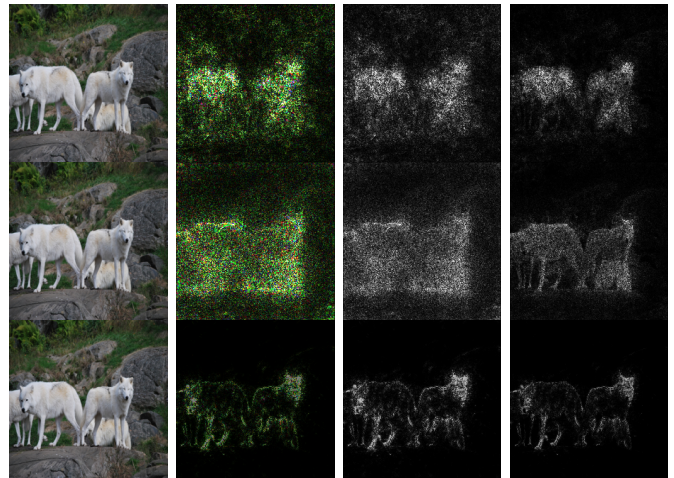

Fig. 5: Example of the relevance maps while filtering and simplifying visualizations. Lines correspond to the explanation method being used, which are the Gradient, SmoothGrad and LRP methods, from top to bottom. The filtering steps are presented in each column: (left) relevance in three channels, (middle) mapping into a single channel, and (right) relevance map multiplied by the original image.

TABLE II: APEM values calculated after each of the three simplification steps. APEM performance decreases as the simplification process proceeds.

	Average			Median		
	1	2	3	1	2	3
Gradient	**231.76**	176.86	127.79	**137**	105	81
Smooth Grad	**93.40**	74.61	44.38	**41**	28	16
LRP	**101.65**	93.93	66.07	**59**	51	41
Guided Bp	**118.65**	103.09	78.07	**63**	53	43
Grad-CAM	**100.86**	100.86	42.25	**41**	41	20
Guided Grad-CAM	**83.00**	69.95	37.79	**51**	42	31

Figure 5 shows examples of relevance maps obtained at each stage for the three explanation methods. Each method behaves differently – some present great changes after each step, and for others only minimal changes are observed.

These filtering steps are likely to discard information that is important to the model. Indeed, APEM values decrease as filtering proceeds. Thus, we evaluate the loss of information that is lost during the filtering steps by using the average and median APEM values. Results are shown in Table II, and they follow the same trend that was observed in Figure 5: explanation methods that drastically change the relevance map during the filtering steps also present the greatest APEM losses, while methods that only produce minimal change are associated with the smaller APEM reductions.

Therefore, the application of these simplifications should be used considering the decrease in APEM performance compared to the amount of visual comprehension that they bring. Explanations which look more *noisy* might be harder for an user to analyze than centered clouds of interpretable information, even if *cleaning* the image means a loss in
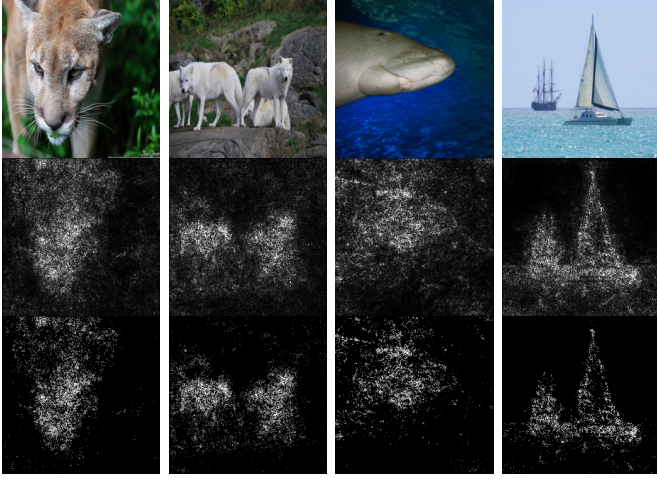
Fig. 6: Examples of the final images obtained with our filtering algorithm. The figure presents input images (top), their relevance maps based on the Gradient method before (middle) and after (bottom) the filtering process.
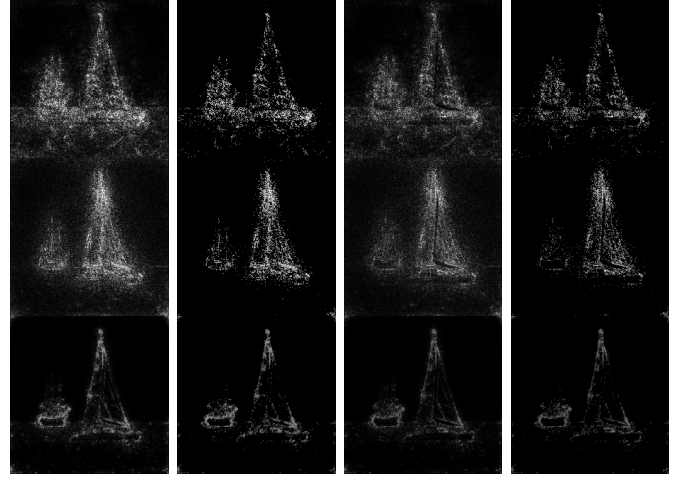

Fig. 7: Filtering algorithm applied to an image's relevance map. The rows correspond to the Gradient, SmoothGrad and LRP methods and the columns are, from left to right: (1) the relevance maps; (2) its filtered image; (3) the map multiplied by the original image; (4) its filtered image.

explanation. These factors should be weighted when applying explanation methods to practical applications.

D. Filtering Explanations

Next we expand the use of APEM to filter relevance maps so that they become more interpretable without any loss of essential information for the model, as we discussed in Section III-B. Specifically, our objective is to make the visualization of relevance maps more interpretable as long as APEM values do not drop. In this case, the relevance maps are made more understandable, while still reliable – in the sense that they comprise the features that actually impact model decision.

Figure 6 shows examples of relevance maps computed for the images and their last configurations before there is a drop in their APEM values. Interestingly, as the filtering steps proceed, regions outside the main object in the image were mostly erased from the relevance map. This means that the small relevance values attributed to the image's context were not actually relevant. For instance, the grass, stones and water present in the examples were given relevance values but they could be removed, leaving only the main objects in the images.

Our filtering algorithm can also be used within other explanation methods, as shown in Figure 7. Each map focuses on slightly distinct parts of the boat because of the characteristics of the explanation method, but all of them removed the relevance that was attributed to the sea, showing that this context was not considered relevant. Finally, the outcome of the filtering algorithm seems easier to interpret than the original noisy maps.

V. FURTHER ANALYSIS

In this section we discuss interesting properties of APEM that could benefit future applications.
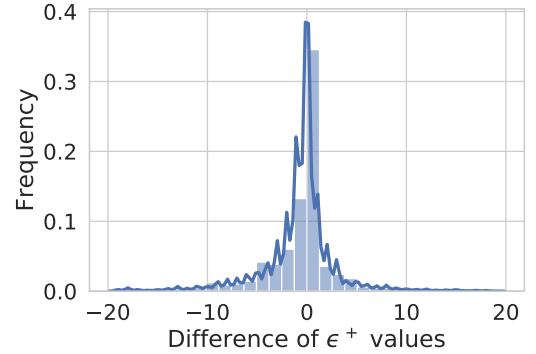
Fig. 8: Histogram and kernel density estimate of the difference between ϵ^+ of LRP and SmoothGrad methods. Positive values are observed when ϵ^+ for LRP is higher than for SmoothGrad.

A. The Importance of Irrelevance Maps

As discussed previously, the inverse of relevance is also taken into account while calculating APEM. This is important because it prevents the explanation methods from focusing on a few relevant pixels, while not giving importance to others that may be also relevant. In this sense, considering irrelevance gives the metric a *recall*-like property. In order to investigate the importance of also considering ϵ^+ (irrelevance) when calculating APEM, we compared LRP with SmoothGrad while only considering irrelevance values. Figure 8 shows the histogram and the kernel density estimate of the difference between ϵ^+ values from LRP and SmoothGrad. Although LRP presents higher APEM values than SmoothGrad, we observed that LRP produces relevance maps that are often more focused than those produced by SmoothGrad. LRP is, then, penalized for this and has lower ϵ^+ values than SmoothGrad on average – a lower ϵ^+ results in a reduction in its APEM value.
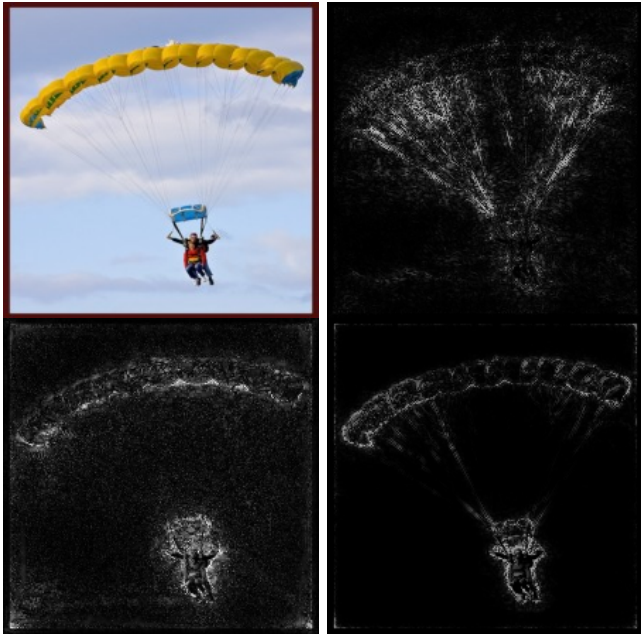
Fig. 9: Example of an instance where the Gradient explanation results in worse APEM results than both LRP and Smooth-Grad. Images represent: (top-left) original image, (top-right) Gradient, (bottom-left) SmoothGrad and (bottom-right) LRP.

B. Single Instance Comparison

We analyze an (outlier) instance where the Gradient results in negative APEM values, which means its resulting irrelevance would be a better predictor of labeling choice than the relevance itself. For this specific image, shown in Figure 9, both LRP and SmoothGrad have better APEM values than the Gradient. In this instance, we can see that the Gradient focuses on the parachute strings and the clouds. Even though the gradient might be locally strong in that region, it is not a good predictor of class change, and, thus, not particularly relevant as an explanation of the model's labeling choice. At the same time, both LRP and SmoothGrad focus on the person and the parachute themselves, resulting in better APEM values.

C. Misclassified Images

So far our analysis only considered a set of $5,000$ correctly classified images. In order to properly explain and debug a model, we also have to understand how APEM behaves with misclassified images. Thus, we performed the same evaluation process on a set of $5,000$ misclassified images. In this case, the labels we used to create relevance maps were the ones predicted by the model.

As (hopefully) expected, the model is more uncertain about a decision when it predicts a wrong label. Then, it is expected that the model does not need many perturbations to make it change a prediction, which results in lower ϵ values. This leads to lower APEM values. On the other hand, this should not influence the quality of the explanation, disregarding the understanding of the generated relevance map. To put it

TABLE III: Average and Median APEM values for correctly classified and misclassified images. Relevance maps are calculated considering the prediction as the ground truth. Misclassification leads to an overall APEM decrease.

	Loss		Median	
	Correct	Misclassified	Correct	Misclassified
Gradient	**127.79**	**43.58**	**81**	**25**
Smooth Grad	44.38	12.65	16	4
LRP	66.07	20.34	41	11

simple, the overall APEM values are decreased but the best explanation methods should to keep their ranking positions.

Table III shows average and median APEM values, comparing them with the ones obtained with the correctly classified images. The APEM decrease is clear, but the relative ordering of the explanation methods remains the same.

D. Correlation between APEM and Loss

The last set of experiments is devoted to investigate the possible correlation between APEM and loss. For this, we used the correctly classified images, misclassified images, and the total set of images. Again, the relevance map calculated for the misclassified images is based on the model prediction even though the loss uses the ground truth. The greatest probability for a label in the prediction is referred to as confidence, and it is also compared with the loss.

We compute the correlation using the Spearman's rank correlation coefficient [37] 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 it is close to -1 when they have a dissimilar one.

Table IV shows the correlations and the statistical significance. Our analysis indicates that correctly classified images have higher correlations while the misclassified images have virtually none. Also, observations have a dissimilar rank between APEM and loss, resulting in a negative correlation. Further, the methods that achieve higher APEM values also are the ones with higher correlation. In summary, the best explanation methods in terms of APEM have a higher correlation with the loss. Finally, high APEM values mean lower losses. Therefore, good explanations given by high APEM values may be used to assess the reliability of the model output.

VI. Conclusions

In this work, we proposed the Adversarial Perturbation Explanation Measure (APEM), a robust measure which evaluates the reliability of explanation methods. APEM enables us to compare explanation methods quantitatively, thus 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 present a comparison of some well-known explanation methods using our proposed measure. Along with it, we also present

TABLE IV: Spearman correlation between APEM values for each explanation method and the loss of the evaluated model († represents statistical significance with $\rho < 0.01$). The correlation of the confidence of the most probable label and the loss is also presented for comparison.

	Loss		
	Correct	Misclassified	Full
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†

some characteristics of the methods and how APEM behaves. Furthermore, we showed some properties which especially make APEM robust. 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 misclassified images, showing that APEM drastically falls when the model is not able to correctly predict an instance. Finally, we correlated APEM results to the output of the model in different situations.

We also studied simplifications that aim to improve visualization of relevance maps. We showed the effect of these simplifications on the reliability of the resulting images, and a simple algorithm that works around this problem. The algorithm is one of the applications in which APEM can be used as a tool to filter the relevance maps into more interpretable images, while all the essential information is kept. The proposed algorithm can be used within an explanation method to create less noisy images and to facilitate its understanding.

REFERENCES

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[2] D. Shen and G. W. H.-I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Eng.*, vol. 19, no. 1, pp. 221–248, 2017.

[3] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A flexible approach for counterfactual prediction," in *Proc. of ICML*, 2017, pp. 1414–1423.

[4] A. Albuquerque, T. Amador, R. Ferreira, A. Veloso, and N. Ziviani, "Learning to rank with deep autoencoder features," in *Proc. of IJCNN*, 2018, pp. 1–8.

[5] T. Alves, A. Laender, A. Veloso, and N. Ziviani, "Dynamic prediction of ICU mortality risk using domain adaptation," in *Proc. of IEEE Big Data*, 2018, pp. 1328–1336.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proc. of ACM SIGKDD*, 2016, pp. 1135–1144.

[7] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K. Müller, "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, pp. 1803–1831, 2010.

[8] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. of IEEE ICCV*, 2017, pp. 3449–3457.

[9] D. Valle, N. Ziviani, and A. Veloso, "Effective fashion retrieval based on semantic compositional networks," in *Proc. of IJCNN*, 2018, pp. 1–8.

[10] G. Cadamuro, R. Gilad-Bachrach, and X. Zhu, "Debugging machine learning models," in *ICML Workshop on Reliable Machine Learning in the Wild*, 2016.

[11] I. Hata, A. Veloso, and N. Ziviani, "Learning accurate and interpretable classifiers using optimal multi-criteria rules," *JIDM*, vol. 4, no. 3, pp. 204–219, 2013.

[12] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. of NeurIPS*, 2018, pp. 9525–9536.

[13] A. Mauricio, F. Cappabianco, A. Veloso, and G. Cámara, "A sequential approach for pain recognition based on facial representations," in *Proc. of ICVS*, 2019, pp. 295–304.

[14] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. of IEEE CVPR*, 2015, pp. 427–436.

[15] A. Marczewski, A. Veloso, and N. Ziviani, "Learning transferable features for speech emotion recognition," in *Proc. ACM Multimedia*, 2017, pp. 529–536.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014.

[17] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, 07 2015.

[18] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 28, no. 11, pp. 2660–2673, 2017.

[19] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Technical Report, Univeristé de Montréal*, 01 2009.

[20] Q. Le, "Building high-level features using large scale unsupervised learning," in *Proc. of IEEE ICASSP*, 2013, pp. 8595–8598.

[21] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. of ECCV*, 2014, pp. 818–833.

[22] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. of IEEE CVPR*, 2015, pp. 5188–5196.

[23] A. Dosovitskiy and T. Brox, "Inverting visual representations with convolutional networks," in *Proc. of IEEE CVPR*, 2016, pp. 4829–4837.

[24] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, vol. abs/1312.6034, 2013.

[25] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017.

[26] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2014.

[27] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. of IEEE ICCV*, 2017, pp. 618–626.

[28] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. of ICML*, 2017, pp. 3319–3328.

[29] A. Binder, S. Bach, G. Montavon, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for deep neural network architectures," in *Proc. of ICISA*, 2016, pp. 913–922.

[30] S. Lapuschkin, A. Binder, G. Montavon, K. Müller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proc. of IEEE CVPR*, 2016, pp. 2912–2920.

[31] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013.

[32] G. Montavon, W. Samek, and K. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *Int. Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

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

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE CVPR*, 2016, pp. 770–778.

[37] C. Croux and C. Dehon, "Influence functions of the spearman and kendall correlation measures," *Statistical Methods and Applications*, vol. 19, no. 4, pp. 497–515, 2010.