AN EVALUATION OF QUALITY AND ROBUSTNESS OF SMOOTHED EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explanation methods play a crucial role in helping to understand the decisions of deep neural networks (DNNs) to develop trust that is critical for the adoption of predictive models. However, explanation methods are easily manipulated through visually imperceptible perturbations that generate misleading explanations. The geometry of the decision surface of the DNNs has been identified as the main cause of this phenomenon and several *smoothing* approaches have been proposed to build more robust explanations. In this work, we provide a thorough evaluation of the quality and robustness of the explanations derived by smoothing approaches. Their different properties are evaluated with extensive experiments, which reveal the settings where the smoothed explanations are better, and also worse than the explanations derived by the common Gradient method. By making the connection with the literature on adversarial attacks, we further show that such smoothed explanations are robust primarily against additive ℓ_p-norm attacks. However, a combination of additive and non-additive attacks can still manipulate these explanations, which reveals shortcomings in their robustness properties.

1 INTRODUCTION

Explanation methods attribute a numerical value to each data feature in order to quantify its relative importance towards the model's prediction. Such attributions help to better understand and trust complex models like deep neural networks (DNNs). In safety-critical tasks, such an understanding is a prerequisite to the deployment of DNNs, because a domain expert will never make important decisions based on a model's prediction unless that model is trustworthy. Moreover, explanations can help to understand the reasons behind the decision of a model, and when it comes to model debugging, they can reveal the presence of any spurious data correlations that may lead to faulty predictions during inference (Ribeiro et al., 2016).

In the context of image classification with deep neural networks, several explanation methods have been proposed based on the gradient with respect to input, also called gradient-based explanations (Baehrens et al., 2010; Bach et al., 2015; Selvaraju et al., 2017; Sundararajan et al., 2017; Springenberg et al., 2015). The explanation generated by these methods, *a saliency map*, highlights the parts of the image that contributed to the prediction. Recent work has shown that gradient-based explanations of neural networks can be fragile and can be easily manipulated via adversarially perturbed inputs (Ghorbani et al., 2019; Dombrowski et al., 2019; Heo et al., 2019; Viering et al., 2019; Kindermans et al., 2019). That is, one can find a small-norm perturbation to be added to an input (often imperceptible), such that the focus of the explanation changes towards irrelevant features while the model's output remains unchanged. This, in turn, can make explanations inappropriate to help end-users gain trust in a model's prediction.

The large curvature of the decision surface of neural networks has been identified as one of the causes of fragility for gradient-based explanations (Ghorbani et al., 2019; Dombrowski et al., 2019; Wang et al., 2020). To make explanations more robust, a class of approaches proposed smoothing the explanation or making the decision surface of neural networks more smooth (Wang et al., 2020; Dombrowski et al., 2019; Ivankay et al., 2020). We refer to these approaches as *smoothing approaches*. It is worth mentioning that similar methods have been proposed in the context of adversarial robustness, with the aim of flattening the decision surface of neural networks in order to reach more robust predictions (Moosavi-Dezfooli et al., 2019; Qin et al., 2019).

Here, we provide a thorough investigation of the explanations derived by smoothing approaches in terms of *explanation quality* and *robustness*. We employ various tests to assess the quality of these explanations. Each test evaluates a desirable property for explanations, such as: sensitivity to changes in the model, fidelity to the predictor function, etc. In terms of robustness, we show that explanations derived by smoothing approaches only provide robustness against additive ℓ_p norm attacks. Specifically, in this work, we show that compared to additive attacks, attacks based on the combination of spatial transformation (Xiao et al., 2018) and/or color transformation (Laidlaw & Feizi, 2019) together with additive perturbations are more effective in manipulating these explanations. Our contributions can be summarized as follows:

- We study the effectiveness of smoothing approaches to achieve robust explanations. We present results on evaluating both the quality and robustness properties of smoothed explanations.
- We assess the quality of smoothed explanations via presenting the results of various quality tests. Our results demonstrate the pros and cons of smoothed explanations with respect to the following quality aspects: sensitivity to model parameters, class discriminativeness, Infidelity, and sparseness.
- We present results for different combination of additive and non-additive attacks, and show that they are able to manipulate explanations derived by smoothing approaches more successfully. Combining different types of perturbations to achieve stronger attacks has been a topic of investigation in the context of adversarial examples (Jordan et al., 2019). To the best of our knowledge, this is the first time such attacks have been used in the context of explanations.

Related works. There have been several works aiming to make explanations more robust. These works mostly focused on either modifying the explanation method itself or modifying the predictor model to achieve robust explanations. Wang et al. (2020) introduced Uniform Gradient, which is similar to Smooth Gradient unless it uses Uniform noise, and showed that it can hardly be manipulated by additive attacks. Dombrowski et al. (2019) proved that a network with soft-plus activations has a more robust Gradient explanation compared to a ReLU network, given that the parameter β of the soft-plus function is chosen to be sufficiently small. Consequently, they proposed the β-smoothing approach in which they substitute the ReLU activations of a trained network by soft-plus functions with a small β parameter. Wang et al. (2020) introduced a regularization term called *Smooth Surface Regularization (SSR)* to the training objective of a DNN. This training objective penalizes the large curvature of a DNN by regularizing the eigenvalue of the input hessian with the maximum absolute value. Moreover, they showed that adversarial training (Madry et al., 2018) also leads to more robust explanations. This fact can also be deduced from the results of (Moosavi-Dezfooli et al., 2019) as they showed that adversarial training leads to a significant decrease in the curvature of the loss surface with respect to inputs. Anders et al. (2020) proposed an attack in which they adversarially manipulate the model instead of the input in order to manipulate the explanation. Then they propose a modification to the existing explanation methods to make them more robust against such manipulated models. Lakkaraju et al. (2020) proposed a framework for generating robust and stable black box explanations based on adversarial training. Chen et al. (2019) introduced a regularization term to the training objective of neural networks to achieve robust Integrated Gradient explanations. Finally, Dombrowski et al. (2020) developed a theoretical framework to derive bounds on the maximum manipubality of explanations and proposed three different techniques to boost the robustness of explanations. In this work, we show that the robustness of smoothed explanations can be affected by employing a combination of additive and non-additive attacks. Furthermore, we present a through evaluation of the different quality aspects of smoothed explanations.

2 BACKGROUND

First, we provide the definition of an explanation map and then briefly describe the explanation methods we used in this paper. Then we continue with introducing the attacks to explanations and the smoothing approaches we are going to study in this paper.

Consider a model $f : \mathbb{R}^d \to \mathbb{R}^K$ which classifies an input $\mathbf{x} \in \mathbb{R}^d$ into one of the K classes. *An explanation map*, denoted by $h_f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^d$, associates a score to each feature of the input

indicating the relevance of that feature towards the model's prediction. For instance, in the context of image classification, saliency maps associate a score to each pixel of the input image resulting in a heatmap that highlights important regions of the image leading to the model prediction. In this work, we focus on the gradient-based explanations and mainly on the Gradient method. Given a model f and an input $\mathbf{x}$, the Gradient explanation is defined as $\nabla_{\mathbf{x}} f(\mathbf{x})$. Since other gradient-based explanation methods make use of the gradients with respect to input, we argue that our results could be extended to those explanation methods as well. We will also consider two smoothed variants, namely Smooth (Smilkov et al., 2017) and Uniform Gradient (Wang et al., 2020) methods.

2.1 ATTACKS TO MANIPULATE EXPLANATIONS

Similarly to common adversarial attacks (Goodfellow et al., 2015; Moosavi-Dezfooli et al., 2016; Szegedy et al., 2014), recent work has shown that explanations can also be manipulated by adding a small and almost imperceptible perturbation to the input (Ghorbani et al., 2019; Dombrowski et al., 2019). We refer to this class of attacks as *explanation attacks*. There have been various formulations for explanation attacks (Ghorbani et al., 2019; Dombrowski et al., 2019). In this work, we will use the formulation introduced by Dombrowski et al. (2019). In this attack, the attacker tries to find a perturbed input for which the explanation is manipulated to be very similar to a given target explanation map while the output of the model remains approximately unchanged. Note that the target map could be any heatmap in general; however, we used the explanation of a target image as a target map in this work. Below, we will give a formal definition of this attack.

Definition 1 (Targeted manipulation attack)**.** *An explanation $h_f(\mathbf{x})$ for model $f(\mathbf{x})$ is vulnerable to attack at input $\mathbf{x}$ if there exist a perturbed input $\mathbf{x}_{adv}$, such that $h_f(\mathbf{x}_{adv})$ is similar to a given target map h^t but the model's output remains unchanged. An attacker finds $\mathbf{x}_{adv}$ by minimizing the following objective function:*

$$\mathcal{L} = \left\| h_f(\mathbf{x}_{adv}) - h^t \right\|^2 + \gamma_1 \left\| f(\mathbf{x}_{adv}) - f(\mathbf{x}) \right\|^2 + \gamma_2 \mathcal{L}_{reg}(\mathbf{x}, \mathbf{x}_{adv}) \tag{1}$$

The first term in (1) ensures the similarity of the manipulated explanation to the target map, the second term ensures the similarity between the model output for the original and perturbed inputs, and the third term regularizes the perturbation to ensure perceptual similarity between the original and perturbed images. Note that $\mathcal{L}_{reg}$ is defined by the attacker according to the type of the perturbation. The relative weighting of the terms in (1) is controlled by the hyper-parameters γ_1 and γ_2.

2.2 TOWARDS ROBUST EXPLANATIONS

Recent works have tried to define the robustness of explanations in terms of the sensitivity of input gradients to changes in the input data (Wang et al., 2020; Dombrowski et al., 2019). Wang et al. (2020) define the robustness of explanations by the Lipschitz continuity coefficient of the input gradients; a smaller coefficient means that the explanation is less sensitive to the changes in the input and hence more robust. In this regard, a class of approaches to generate robust explanations have been proposed in the recent works, which are either based on smoothing out the explanation maps or flattening the decision boundary of the model itself. Broadly, these approaches can be classified into two categories: (1) *Post-hoc* approaches do not require retraining of the network and can be applied as a post-processing step. (2) *Ad-hoc* approaches to robust explanations require retraining of the network and hence are more costly.

In this work, we consider Smooth Gradient (Smilkov et al., 2017), Uniform Gradient (Wang et al., 2020), and β-smoothing (Dombrowski et al., 2019) as post-hoc approaches. The first two methods involve smoothing the explanation map, while the third one smooths the decision surface of the model. All three approaches act on pre-trained models, and hence are characterized as post-hoc. Among the ad-hoc methods, we study the explanations generated by adversarially trained networks, and networks trained with curvature regularization (CURE) (Moosavi-Dezfooli et al., 2019), which is a similar approach to SSR (Wang et al., 2020)[1].

[1]We experiment only with CURE, because with the publicly available code of SSR we were not able to reproduce the results in (Wang et al., 2020).

3 EVALUATING POST-HOC APPROACHES

Here, we begin by evaluating the *quality* of explanations derived by post-hoc approaches that do not require retraining of the network. Then, we evaluate the *robustness* of these explanations by presenting results on effective non-additive attacks to manipulate them. For all of the experiments in this section, we used a VGG-16 network trained on ImageNet (Russakovsky et al., 2015), and for generating the explanation maps we used the Captum (Kokhlikyan et al., 2020) package. Moreover, for the β-smoothing approach we always set $\beta = 0.8$ as suggested in (Dombrowski et al., 2019).

3.1 QUALITY OF EXPLANATIONS OF POST-HOC APPROACHES

To evaluate and compare the quality of the explanations, we use various quality tests presented in the literature. In general, assessing the quality of an explanation is a challenging task and each quality test only evaluates a specific quality aspect of an explanation. Therefore the assemblage of quality tests helps to understand which quality aspects of the explanations are improved and which are deteriorated by the smoothing approaches.

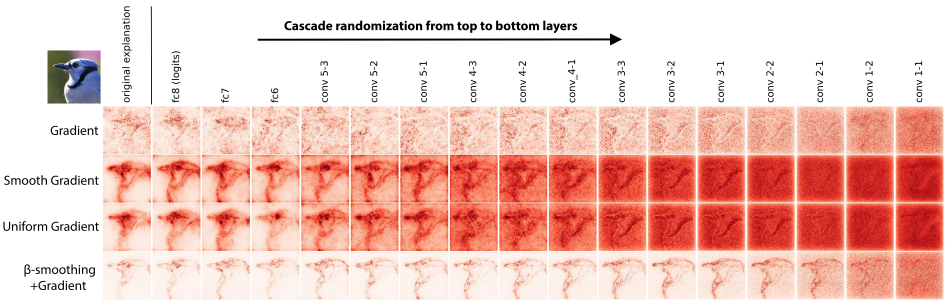

Figure 1: Cascade randomization of the VGG-16 (ImageNet) layers. The first column shows the original explanation for the image "Jay" bird. Each subsequent column shows the effect of randomization of the parameters of the network up to that layer (inclusive) on the explanations.

Cascade randomization of model parameters. Adebayo et al. (2018) argued that it is desired for an explanation to be sensitive to the changes in the model parameters. They proposed a model parameter randomization test to assess this sensitivity. In this test, the parameters of a model are progressively randomized from the top layer (logits) to the bottom layers. In each step of randomization, the explanation from the resulting model is compared against the explanation from the original model. Randomizing the model parameters means losing what the model has learned from the data during training. Therefore, we expect a "good" explanation to be destroyed in this process. However, if an explanation is insensitive to the randomization of the model parameters, then it is not deemed appropriate for debugging the model under erroneous predictions.

The visual results of this test for Gradient explanation and post-hoc approaches are shown in Figure 1. More examples of this test can be found in the Appendix. One can observe that the explanations derived from post-hoc approaches show less sensitivity to the randomization of model parameters than compared to the Gradient method. This can also be verified by the Spearman rank correlation between the original and randomized explanations shown in Figure 2. We observe that for the smoothed explanation methods, the original and randomized explanations have a high rank correlation after the randomization of the top layers of the network. *These results highlight that using Smooth Gradient, Uniform Gradient, and β-smoothing to achieve a more robust explanation can come at the expense of having explanations that are less sensitive to model parameters.*

Class sensitivity of explanations. A good visual explanation should be able to localize the image regions relevant to the target category, i.e., it should be *class discriminative* (Selvaraju et al., 2017). This is particularly significant when dealing with images containing more than one object. To assess the class discriminativeness of an explanation we used a quality test equivalent to the pointing game (Zhang et al., 2016). We sampled images from the MS COCO dataset (Lin et al., 2014), containing two objects that are also present among the ImageNet class labels. For this test we only keep

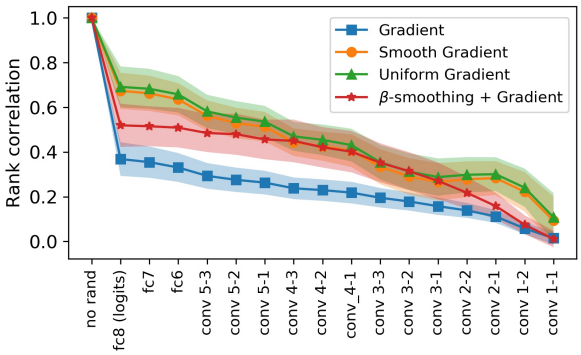

Figure 2: Spearman rank correlation between the original and randomized explanation derived for randomization up to the layer indicated by the x-axis. A higher rank correlation value indicates a higher similarity, i.e, *the higher the curve of an explanation method, the less sensitive it is to model parameters.* The results are averaged over 1000 Images from Imagenet and the shaded area around each curve indicates the standard deviation.

Table 1: Ratio of the explanation top-20 values included in the segmentation mask of each object. Note that the columns "obj 1" and "obj 2" are for the explanations computed for the top predicted class and the class corresponding to the second object in the image respectively. The results are averaged over 60 samples.

Explanation Method	obj 1	obj 2
Gradient	0.6	**0.49**
Smooth Gradient	0.54	0.37
Uniform Gradient	0.57	0.38
β-smoothing + Gradient	**0.62**	0.45

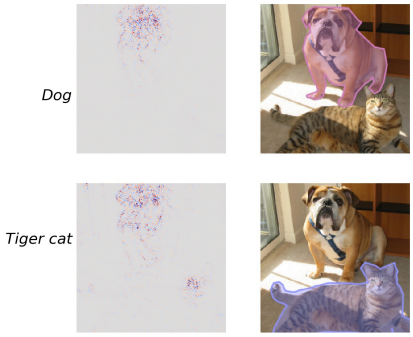

Figure 3: An example of β-smoothing explanation generated for the target category "dog" and "cat" (left), and the segmentation mask of each object from the COCO dataset (right).

the samples for which one of the objects in the image is the top predicted class by the network and the other object is among the top 20 predicted classes by the network. We compute the explanation maps for each of the class labels corresponding to the objects. Using the segmentation mask of the objects provided in the dataset as ground truth, we compute what percentage of the top-20 values in the explanation maps generated for each target category are inside the corresponding segmentation masks. The results of this test are shown in table 1 and a visual depiction of this test is given in Figure 3. These results indicate that the smoothed explanation methods are less discriminatory when generated for the target class label that has a lower probability. *This suggests that in terms of class discriminativeness of explanations, the post-hoc smoothing approaches investigated in this paper are inferior to the Gradient method.*

Sparseness of explanations. To create explanations that are human-accessible, it is advantageous to have a *sparse* explanation map (Molnar, 2019), i.e, only the features that are truly predictive of the model output should have significant contributions, and irrelevant features should have negligible contributions. Sparse explanations are more concise because they only include features with significant contribution making it simpler for end-users to understand the reasons for a specific prediction of the model (Chalasani et al., 2020). To measure the sparseness of an explanation map, we applied the Gini Index on the absolute value of the flattened explanation maps. The Gini Index is a metric that measures the sparseness of a vector with non-negative values (Hurley & Rickard, 2009). By definition, the Gini Index take values in $[0, 1]$ with higher values indicating more sparseness. Table 2 shows the average Gini Index of the Gradient, Smooth Gradient, Uniform Gradient, and

Table 2: Average Gini Index for the explanations of a VGG-16 network (averaged over 1000 samples). A **larger** Gini Index indicates more sparseness and hence a more concise explanation.

	Gradient	Smooth Gradient	Uniform Gradient	β-smoothing
Gini Index	0.56 ± 0.038	0.34 ± 0.053	0.35 ± 0.058	$\mathbf{0.65} \pm 0.067$

Table 3: Average Infidelity for the explanations of a VGG-16 network (averaged over 100 samples). A **lower** Infidelity value indicates better fidelity of the explanation to the predictor function.

	Gradient	Smooth Gradient	Uniform Gradient	β-smoothing
Infidelity	1.43 ± 1.52	1.42 ± 1.52	1.42 ± 1.52	$\mathbf{1.00} \pm 0.88$

β-smoothing computed for 1000 randomly sampled images from ImageNet. *The results show that compared to the Gradient method, Smooth Gradient and Uniform Gradient provide less concise explanations, whereas β-smoothing actually improves the sparseness of the explanations as compared to the Gradient method.*

Explanation Infidelity. Introduced in Yeh et al. (2019), this metric captures how the predictor function changes in response to significant perturbations to the input and is defined as the expected difference between the two terms: 1) the dot product of the input perturbation and the explanation and 2) the difference between function values after significant perturbations to the input. The metric generalizes the completeness axiom (Shrikumar et al., 2017; Sundararajan et al., 2017) because it allows for different types of perturbations which could be of interest depending on the problem and the dataset. We use the infidelity metric to compare the effect of post-hoc smoothing approaches on the fidelity of explanations to the predictor function. As suggested in (Yeh et al., 2019), we used the square removal perturbation to compute the infidelity of explanations for randomly selected images from ImageNet. Table 3 shows the results for the post-hoc approaches. A lower infidelity value indicates better fidelity of the explanation to the predictor function. *The results suggest that the degree of smoothing used to robustify explanations, also improves their infidelity.* Therefore with respect to the Infidelity metric, all of the smoothed explanations investigated in this section are superior to the Gradient method. This finding is also in line with the results of Yeh et al. (2019), i.e., that modest smoothing improves the infidelity of explanations.

3.2 Robustness of explanations of post-hoc approaches

Now, we will evaluate the robustness of Smooth Gradient, Uniform Gradient, and β-smoothing explanations. We present attacks composed of additive and non-additive perturbations, and show that they are more effective than additive attacks to manipulate explanations. The non-additive attacks we employed are spatial transformation attacks (Xiao et al., 2018), and recoloring attacks (Laidlaw & Feizi, 2019). See the Appendix B for a brief description of each of these attacks. In the rest of this paper, we refer to the additive attack as *Delta*, spatial transformation attack as *StAdv*, and recoloring attack as *Recolor*.

We used the projected gradient descent (PGD) algorithm to optimize the objective function (1)[2]. In our experiments, we evaluate three combinations of attacks, namely Delta, Delta+StAdv, and Delta+StAdv+Recolor, against the explanation of a VGG-16 network trained on ImageNet (Russakovsky et al., 2015). See Appendix C.1 for the details about the ℓ_∞ norm for each type of the perturbations and the hyper-parameters used in each attack setting.

We use two metrics to evaluate the attacks: (1) The Cosine Distance metric (cosd) to evaluate the similarity between the target and manipulated explanations (Wang et al., 2020). A lower cosine distance corresponds to a lower ℓ_2 distance between the target and manipulated explanations indicating a higher similarity. The range of the values for cosd is between 0 and 1. (2) The LPIPS metric for

[2]As discussed in (Ghorbani et al., 2019; Dombrowski et al., 2019), to avoid zero-valued gradients when optimizing (1), we have to replace the ReLU activation with its smooth approximation. In this work, we used a soft-plus function with $\beta = 100$.

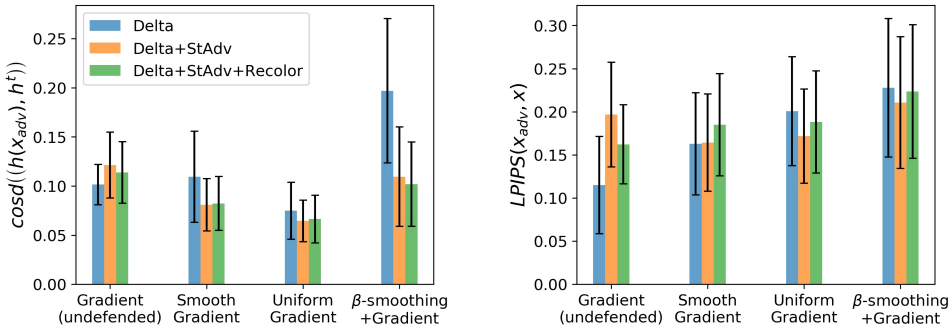

Figure 4: Evaluation with cosd (left) and LPIPS metrics for attacks against post-hoc approaches. For both cosd and LPIPS, smaller values indicate higher similarity. All results are averaged over 200 images from ImageNet. The black bars indicate standard deviation.

quantifying the *perceptual* similarity between images (Zhang et al., 2018). A lower LPIPS value indicates higher similarity.

Figure 4 shows the cosine distance between the target and manipulated explanations, and the perceptual similarity (LPIPS) between the perturbed and original images for each attack setting. We can observe that Delta+StAdv, and Delta+StAdv+Recolor attacks are more effective than Delta attacks to manipulate β-smoothing explanations, i.e, with a less perceptible perturbation (lower LPIPS value), we can reach a cosd value between manipulated and target explanations very close to the cosd value when attacking the Gradient method. The effect of the non-additive attacks is less significant on the Smooth and Uniform Gradient methods, however we can still observe improvements in the cosd values under these attacks. Taken together, these results show that *Smooth Gradient, Uniform Gradient, and β-smoothing explanations are more vulnurale to non-additive attacks and hence such attacks should be considered as a threat to the robustness of these methods*. As an example, we can visually see the effectiveness of Delta+StAdv+Recolor attack against differnt explanation methods in Figure 5.

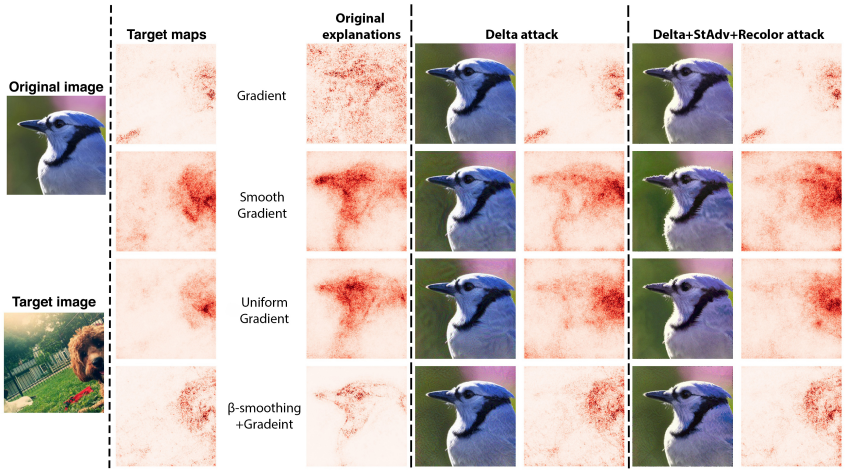

Figure 5: An example of visual comparisons of explanation attacks against different methods. Under each attack setting, the left column shows the perturbed image and the right column shows the corresponding perturbed explanation.

4 Evaluating ad-hoc approaches

Here, we recreate the experiments of Section 3 for the ad-hoc approaches. We study the explanations of networks trained with curvature regularization (CURE) (Moosavi-Dezfooli et al., 2019),

and adversarial training (Madry et al., 2018). Training with CURE, regularizes the eigenvalue of the input hessian with maximum absolute value and is similar to SSR, which was shown to improve the robustness of explanations against additive attacks (Wang et al., 2020). Adversarial training also smooths the decision surface and can provide more robust explanations.

For the experiments in this section, we used a ResNet-18 network trained with CURE and an adversarially trained ResNet-18 network trained on adversarial examples with ℓ_∞ norm of the perturbations upper bounded by $8/255$ (Engstrom et al., 2019). Both networks are trained on CIFAR-10 dataset (Krizhevsky, 2012).

4.1 Quality of explanations of ad-hoc approaches

Cascade randomization of model parameters. We evaluate the sensitivity of explanations of the networks trained with CURE and adversarial training using the cascade randomization of model parameters test.

The Spearman rank correlation between the original and randomized explanations is shown in Figure 6. These Results show the explanation of an adversarially trained network is less sensitive to model parameters. *This suggests that the explanation of an adversarially trained network cannot be helpful to debug a model when it is making a wrong prediction.*

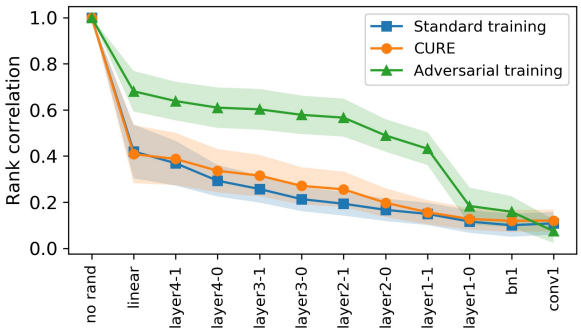

Figure 6: Spearman rank correlation between the original and the randomized explanations (averaged over 1000 Images from CIFAR-10). A higher rank correlation value indicates a higher similarity, i.e, *the higher the curve of an explanation method, the less sensitive it is to model parameters.*

Sparseness of explanations. We compare the sparseness of the explanations derived by ad-hoc approaches, using the Gini Index metric. Table 4 compares the Gini Index for the explanations of networks trained with different training objectives. These results show that adversarial training helps to improve the sparseness of explanations as compared to standard training. Hence the explanations of an adversarially trained network are more *concise*. This is in line with the results of Chalasani et al. (2020) as well. However, the rsults of Table 4 indicates that training a network with CURE does not help to improve the sparseness of explanations as compared to standard training.

Table 4: Gini Index for the explanations of a ResNet-18 network (averaged over 1000 images from CIFAR-10). A **larger** Gini Index suggests a more concise explanation.

	Standard training	CURE	Adversarial training
Gini Index	0.54 ± 0.035	0.54 ± 0.045	$\mathbf{0.71} \pm 0.054$

Explanation Infidelity. To compare the fidelity of explanations derived by ad-hoc approaches to the predictor function, we used the Infidelity metric with square perturbation (Yeh et al., 2019). Table 5 shows the results for randomly selected images from CIFAR-10. A lower infidelity value indicates better fidelity of the explanation to the predictor function. From these results, we can observe that training a network with CURE and adversarial training helps to improve the explanation Infidelity. Therefore with respect to the Infidelity metric, the ad-hoc smoothing approaches investigated in this section improve the explanation Infidelity as compared to standard training.

Table 5: Infidelity of the explanations of a ResNet-18 network (averaged over 1000 images from CIFAR-10). A **lower** Infidelity value is better.

	Standard training	CURE	Adversarial training
Infidelity	5.69 ± 4.44	$\mathbf{0.59} \pm 0.34$	1.56 ± 0.76

4.2 Robustness of explanations of ad-hoc approaches

Now, we evaluate the improvement of robustness via ad-hoc approaches. We present results for Delta, Delta+StAdv, and Delta+StAdv+Recolor attacks against explanations of the networks trained with CURE and adversarial training. Figure 7 shows the results of these attacks. For the adversarially trained network, we can observe that non-additive attacks can more effectively manipulate explanations compared to the additive attacks. However, even with the strongest attack setting we still cannot get close to the cosd value reached by attacking the explanation of the network trained in standard way. For the attacks against the explanation of the network trained with CURE, the effect of non-additive attacks are less significant in terms of the cosd value, however we can still observe that such attacks can reach similar cosd values with perceptually less visible perturbations.

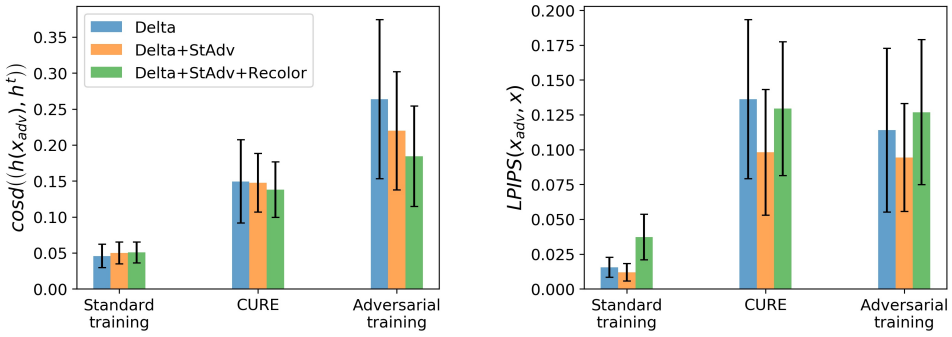

Figure 7: Evaluation with cosd (left) and LPIPS (right) metrics for the attacks against ad-hoc approaches. For both cosd and LPIPS, smaller values indicate higher similarity. All results are averaged over 1024 images from CIFAR-10. The back bars indicate standard deviation.

5 Conclusion

We have evaluated two aspects of smoothed explanations: a) explanation quality, and b) robustness of explanation. In terms of explanation quality, we performed a thorough evaluation of four quality aspects: sensitivity to model parameters, class discriminativeness, sparseness, and infidelity. Our results show that the smoothed explanations investigated in this paper perform worse than those of the Gradient method in terms of sensitivity to model parameters and class discriminativeness. On the other hand, we show that using such smoothing methods helps to improve explanation Infidelity and sparseness.

We further looked at the robustness of explanations, when inputs are perturbed by a combination of additive and non-additive attacks. To the best of our knowledge, this is the first time such attacks are used to manipulate explanations. Our experimental results highlighted the fact that non-additive attacks are still a threat for explanation methods, including the smoothed ones. These results also point us to the fact that many problems in explanation robustness can be addressed by making analogies with the area of prediction robustness. As these two areas are closely related, the solutions already explored in prediction robustness can be potentially helpful to study explanation robustness. This will be the focus of our future work.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò

Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9525–9536, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Christopher J. Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 314–323. PMLR, 2020. URL http://proceedings.mlr.press/v119/anders20a.html.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, 2010. URL http://portal.acm.org/citation.cfm?id=1859912.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1383–1391. PMLR, 2020. URL http://proceedings.mlr.press/v119/chalasani20a.html.

Jiefeng Chen, Xi Wu, Vaibhav Rastogi, Yingyu Liang, and Somesh Jha. Robust attribution regularization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14300–14310, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/172ef5a94b4dd0aa120c6878fc29f70c-Abstract.html.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 13567–13578, 2019. URL http://papers.nips.cc/paper/9511-explanations-can-be-manipulated-and-geometry-is-to-blame.

Ann-Kathrin Dombrowski, Christopher J. Anders, Klaus-Robert Müller, and Pan Kessel. Towards Robust Explanations for Deep Neural Networks. *arXiv e-prints*, art. arXiv:2012.10425, December 2020.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 3681–3688. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://doi.org/10.1609/aaai.v33i01.33013681.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2921–2932, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html.

Niall P. Hurley and Scott T. Rickard. Comparing measures of sparsity. *IEEE Trans. Inf. Theory*, 55(10):4723–4741, 2009. doi: 10.1109/TIT.2009.2027527. URL https://doi.org/10.1109/TIT.2009.2027527.

Adam Ivankay, Ivan Girardi, Chiara Marchiori, and Pascal Frossard. FAR: A general framework for attributional robustness. *CoRR*, abs/2010.07393, 2020. URL https://arxiv.org/abs/2010.07393.

Matt Jordan, Naren Manoj, Surbhi Goel, and Alexandros G. Dimakis. Quantifying perceptual distortion of adversarial examples. *CoRR*, abs/1902.08265, 2019. URL http://arxiv.org/abs/1902.08265.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pp. 267–280. Springer, 2019. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10408–10418, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6e923226e43cd6fac7cfe1e13ad000ac-Abstract.html.

Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5628–5638. PMLR, 2020. URL http://proceedings.mlr.press/v119/lakkaraju20a.html.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv e-prints*, art. arXiv:1405.0312, May 2014.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Christoph Molnar. *Interpretable Machine Learning*. 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2574–2582. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.282. URL https://doi.org/10.1109/CVPR.2016.282.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 9078–9086. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00929. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Moosavi-Dezfooli_Robustness_via_Curvature_Regularization_and_Vice_Versa_CVPR_2019_paper.html.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13824–13833, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0defd533d51ed0a10c5c9dbf93ee78a5-Abstract.html.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. In *Proceedings of the Eleventh Annual International Conference of the Center for Nonlinear Studies on Experimental Mathematics: Computational Issues in Nonlinear Science: Computational Issues in Nonlinear Science*, pp. 259–268, USA, 1992. Elsevier North-Holland, Inc.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 618–626. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *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*, pp. 3145–3153. PMLR, 2017. URL http://proceedings.mlr.press/v70/shrikumar17a.html.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6034.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6806.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *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*, pp. 3319–3328. PMLR, 2017. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Tom J. Viering, Ziqi Wang, Marco Loog, and Elmar Eisemann. How to manipulate cnns to make them lie: the gradcam case. *CoRR*, abs/1907.10901, 2019. URL http://arxiv.org/abs/1907.10901.

Zifan Wang, Haofan Wang, Shakul Ramkumar, Piotr Mardziel, Matt Fredrikson, and Anupam Datta. Smoothed geometry for robust attribution. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/9d94c8981a48d12adfeecfe1ae6e0ec1-Abstract.html.

Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HyydRMZC-.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10965–10976, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html.

Jianming Zhang, Zhe L. Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pp. 543–559. Springer, 2016. doi: 10.1007/978-3-319-46493-0_33. URL https://doi.org/10.1007/978-3-319-46493-0_33.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

A FORMAL DEFINITION OF EXPLANATION METHODS

Definition 2 (Gradient (Baehrens et al., 2010; Simonyan et al., 2014)). *Given a model $f(\mathbf{x})$, the* gradient *explanation for an input $\mathbf{x}$ is defined as: $\nabla_{\mathbf{x}} f(\mathbf{x})$.*

Definition 3 (Smooth Gradient (Smilkov et al., 2017)). *Given a model $f(\mathbf{x})$ and a user-defined variance σ^2, the* smooth gradient *explanation is defined as: $\mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{x},\sigma^2 I)}\, \nabla_{\mathbf{z}} f(\mathbf{z})$.*

Definition 4 (Uniform Gradient (Wang et al., 2020)). *Given a model $f(\mathbf{x})$ and a user defined standard deviation σ, the* uniform gradient *map is defined as: $\mathbb{E}_{\mathbf{u}\sim\mathcal{U}(\mathbf{x}-\sigma,\mathbf{x}+\sigma)}\, \nabla_{\mathbf{u}} f(\mathbf{u})$, i.e, it is computed by taking average of the samples of $\mathbf{x}$ which are perturbed by adding Uniform noise.*

B DESCRIPTION OF ATTACKS

Additive attack. In this attack we try to find a perturbation δ such that the perturbed input $\mathbf{x}_{adv} = \mathbf{x} + \delta$ can successfully manipulate the explanation according to the Definition 1. The similarity of $\mathbf{x}_{adv}$ to $\mathbf{x}$ is ensured by imposing a constraint on the ℓ_∞ norm of the perturbation, i.e, $\|\delta\|_\infty < \epsilon_{\text{additive}}$. Note that using this attack, there is no further need to have the third regularization term in (1). This is the attack which have been used in most of the works in the context of explanations attacks (Ghorbani et al., 2019; Dombrowski et al., 2019; Wang et al., 2020).

Spatial transformation attack. In this attack we try to find a flow field f which independently displaces each pixel of the input image $\mathbf{x}$ in order to reach a perturbed image $\mathbf{x}_{adv}$ (Xiao et al., 2018). In order to ensure a high similarity between $\mathbf{x}$ and $\mathbf{x}_{adv}$, we impose a constraint on the ℓ_∞ norm of the flow, i.e, $\|f\|_\infty < \epsilon_{spatial}$. This means that we want to restrict the maximal displacement of a pixel. Moreover, we regularize the total variation of the flow (Rudin et al., 1992; Xiao et al., 2018) to ensure a locally smooth spatial transformation. This regularization term would substitute the third term in (1) in case of spatial transformation attack.

Recoloring attack. This class of attacks were introduced in (Laidlaw & Feizi, 2019) as functional adversarial attacks. In this attack the adversarial image $\mathbf{x}_{adv}$ is derived by applying a single function F to each pixel of the input image $\mathbf{x}$. The function F is to be learned during the optimization. In other words, this is equivalent to uniformly changing the colors of the input image, and hence this attack could also be considered as recoloring the input image. Formally, consider a given pixel $\mathbf{x}_i = (c_{i,1}, c_{i,2}, c_{i,3}) \in \mathcal{C}$ in the input image $\mathbf{x}$ where $\mathcal{C}$ is a color space, e.g, RGB. Each pixel in the perturbed image $\mathbf{x}_{adv}$ is derived by applying the function $F(.)$ to the color in the corresponding pixel in $\mathbf{x}$. i.e,

$$\mathbf{x}_{adv,i} = (c_{adv,i,1}, c_{adv,i,2}, c_{adv,i,3}) = F\left(c_{i,1}, c_{i,2}, c_{i,3}\right)$$

To ensure that no pixel can be perturbed by more than a certain amount along each dimension in the color space we impose the condition $|c_i - c_{adv,i}| < \epsilon_i, i = 1, 2, 3$ to the learned function F. Also to ensure that similar colors are perturbed similarly, we use a regularization term based on total variation of the color differences caused by F which substitutes the third term in (1) in case of recoloring attack.

C SETTINGS OF THE ATTACKS

C.1 HYPER-PARAMETERS OF THE ATTACKS TO THE EXPLANATIONS DERIVED BY POST-HOC APPROACHES

In the implementation of the targeted manipulation attack, in addition to γ_1, and γ_2 in (1), we also used another hyper-parameter γ_0 in order to control the relative weighting of the first term in (1). Therefore, γ_0, and γ_1 control the weighting of the mean squared error of the explanations and the network outputs respectively. γ_2 controls the weighting of the smooth loss in case of spatial transformation or recoloring attacks. Note that for evaluating post-hoc approaches we used models trained on Imagenet dataset. Therefore, for tuning the hyper-parameters we started from the values suggested in (Dombrowski et al., 2019) for Imagenet models and updated those values for our attack settings and explanation methods. Table 6 shows a summary of the hyper-parameters for each combination of attack type and explanation method.

Table 6: Hyper-parameters used in our attacks to explanations derived by post-hoc approaches

Method	Attack type	lr	iterations	γ_0, γ_1, γ_2
(Smooth/Uniform) Gradient	Delta	0.001	1500	10^{11}, 10^5, -
(Smooth/Uniform) Gradient	Delta+StAdv	0.0004	1500	10^{11}, 10^5, 0.05
(Smooth/Uniform) Gradient	Delta+StAdv+Recolor	0.0003	1500	10^{11}, 10^5, 0.05
β-smoothing+Gradient	Delta	0.00025	500	10^{11}, 10^5, -
β-smoothing+Gradient	Delta+StAdv	0.0002	500	10^{11}, 10^5, 0.05
β-smoothing+Gradient	Delta+StAdv+Recolor	0.0002	500	10^{11}, 10^5, 0.05

Table 7: Hyper-parameters used in our attacks to explanations derived by ad-hoc approaches

Method	Attack type	lr	iterations	γ_0, γ_1, γ_2
CURE	Delta	0.0002	1500	10^7, 10^2, -
CURE	Delta+StAdv	0.00015	1500	10^7, 10^2, 0.05
CURE	Delta+StAdv+Recolor	0.0001	1500	10^7, 10^2, 0.05
Adversarial training	Delta	0.0002	1500	10^7, 10^2, -
Adversarial training	Delta+StAdv	0.00015	1500	10^7, 10^2, 0.05
Adversarial training	Delta+StAdv+Recolor	0.0001	1500	10^7, 10^2, 0.05

C.2 Hyper-parameters of the attacks to the explanations derived by ad-hoc approaches

In this work, in order to evaluate ad-hoc approaches we used networks trained on CIFAR-10 dataset. Therefore, for tuning the hyper-parameters we started from the values suggested in (Dombrowski et al., 2019) for CIFAR-10 models and updated those values for our settings. Table 7 shows a summary of the hyper-parameters of the attacks to the explanations derived by ad-hoc approaches.

C.3 Settings of the perturbations.

Additive perturbation. For attacking post-hoc approaches, we set $\epsilon_{additive} = 8/255$ in all attack settings. For attacking ad-hoc approaches (networks trained with CURE or adversarial training on CIFAR-10), we set $\epsilon_{additive} = 12/255$ for the Delta attack, and $\epsilon_{additive} = 4/255$ for the Delta+StAdv, and Delta+StAdv+Recolor attacks.

Spatial transformation. For all of the attacks involving spatial transformation we set $\epsilon_{spatial} = 0.05$.

Recoloring. For recoloring attacks, we use the settings suggested by (Laidlaw & Feizi, 2019). We change the color of the pixels in CIELUV color space and set $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0.06$. The resolution of the grid over which the function $F(.)$ is parametrized is set to $16 \times 32 \times 32$ for CIELUV color space.

D Definition of the Gini Index

The Gini Index is defined in (Hurley & Rickard, 2009). Given a vector of non-negative values $\mathbf{v} = [v_1, v_2, \ldots, v_d]$, suppose that the vector is sorted in non-decreasing order so that the resulting indices after sorting are $(1), (2), \ldots (d)$ i.e, $v_{(k)}$ denotes the kth value in the sorted vector. The Gini Index is given by:

$$G(\mathbf{v}) = 1 - 2\Sigma_{k=1}^{d} \frac{v_{(k)}}{\|\mathbf{v}\|_1} \left(\frac{d - k + 0.5}{d} \right) \tag{2}$$

The Gini Index by definition lies in $[0, 1]$, and a higher value indicates more sparseness. In the extreme case where only one of the $v_i > 0$ and all the rest are 0, the Gini Index is equal to 1 (perfect

sparseness). At the other extreme, if all v_i are equal to some positive constant, the Gini Index is equal to zero.

E DEFINITION OF INFIDELITY

The definition is adopted from (Yeh et al., 2019).

Definition 5. *Given a black-box function f, explanation functional h, a random variable $\mathbf{I} \in \mathbb{R}^d$ with probability measure $\mu_{\mathbf{I}}$, which represents meaningful perturbation of interest, the explanation infidelity of h is defined as:*

$$\text{INFD}(h_f, f, \mathbf{x}) = \mathbb{E}_{\mathbf{I}\sim\mu_{\mathbf{I}}}\left[\left(\mathbf{I}^T h(f, \mathbf{x}) - (f(\mathbf{x}) - f(x - \mathbf{I}))\right)^2\right] \tag{3}$$

where $\mathbf{I}$ represents significant perturbations around $\mathbf{x}$ and can be specified in various ways.

F ADDITIONAL EXPERIMENT ON THE INFIDELITY OF POST-HOC APPROACHES.

Table 8 shows the infidelity of the post-hoc approaches for the explanations of a VGG-16 network trained on the CIFAR-10 dataset. Compared to the Imagenet samples, the Infidelity results on the CIFAR-10 samples shows more significant difference between the Gradient method and post-hoc approaches.

Table 8: Average Infidelity for the explanations of a VGG-16 network trained on the CIFAR-10 dataset (averaged over 200 samples.). A **lower** Infidelity value indicates better fidelity of the explanation to the predictor function.

	Gradient	Smooth Gradient	Uniform Gradient	β-smoothing
Infidelity	30.18 ± 28.56	21.54 ± 26.39	20.95 ± 25.38	$\mathbf{16.46} \pm 21.12$

G ADDITIONAL EXAMPLES OF SANITY CHECKS FOR POST-HOC APPROACHES

Cascade randomization of network layers. In this part we provide additional examples of the cascade randomization test to evaluate the sensitivity of Smooth Gradient, Uniform Gradient, and β-smoothing explanations to the changes in model parameters. For the sake of comparison, we also include the results of this test for the Gradient explanation. Figures 8-12 show the results of this test for more examples.

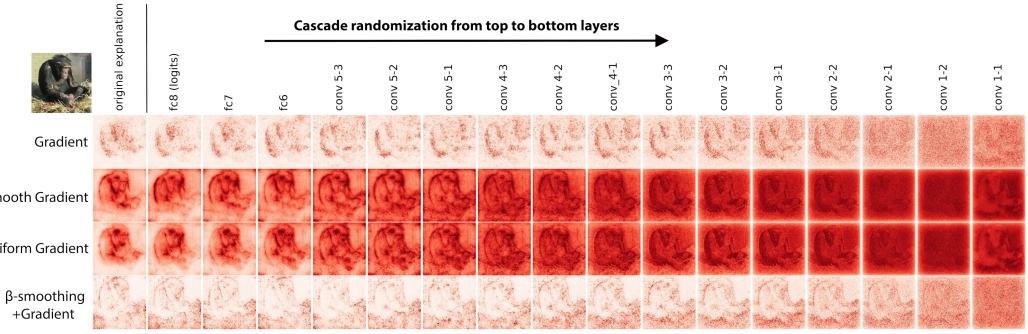

Figure 8: Explanation maps at each step of cascade randomization of the VGG-16 layers.

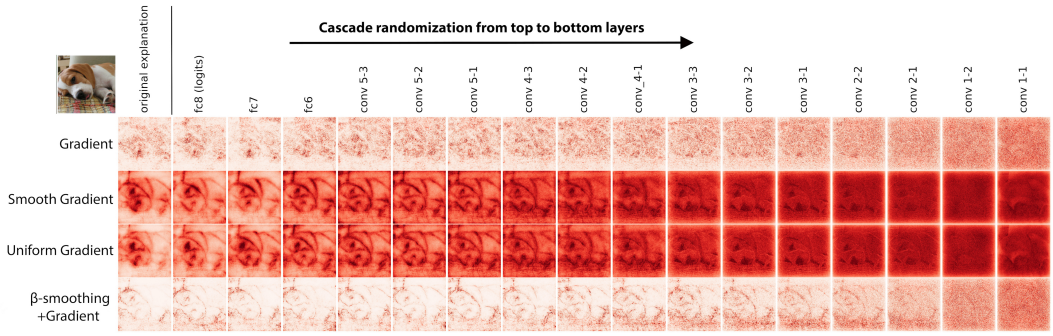

Figure 9: Explanation maps at each step of cascade randomization of the VGG-16 layers.

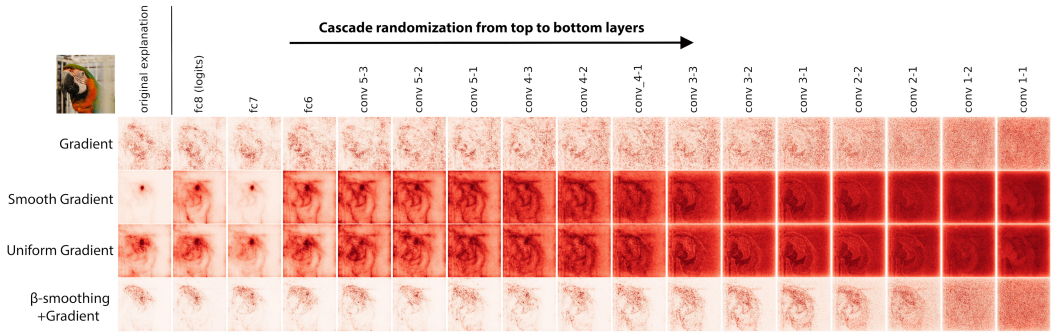

Figure 10: Explanation maps at each step of cascade randomization of the VGG-16 layers.

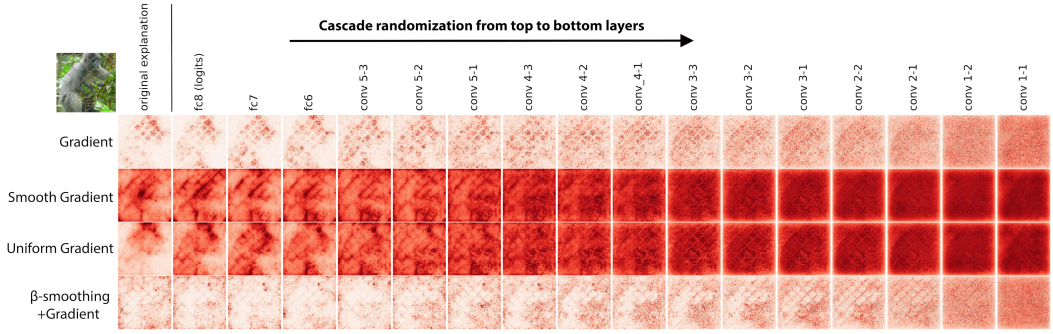

Figure 11: Explanation maps at each step of cascade randomization of the VGG-16 layers.

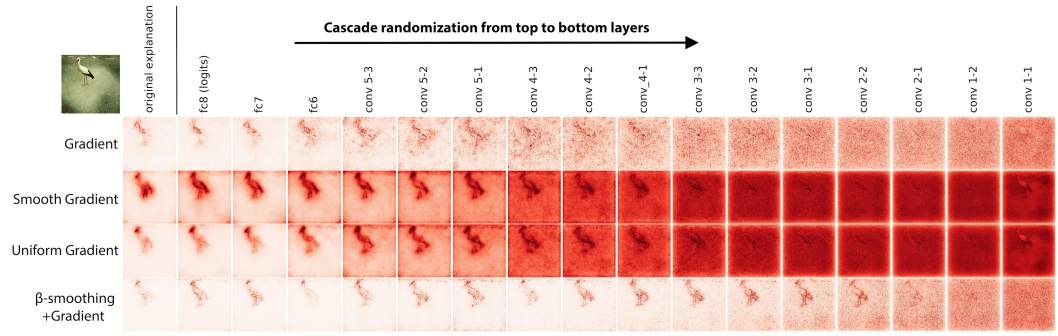

Figure 12: Explanation maps at each step of cascade randomization of the VGG-16 layers.

Class sensitivity of explanations. In this part we provide more visual examples of the pointing game experiment performed to evaluate the class discriminativeness of explanations.

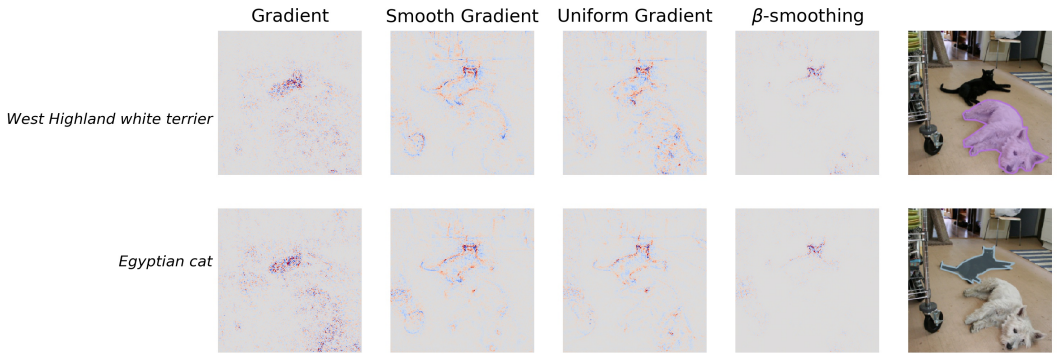

Figure 13

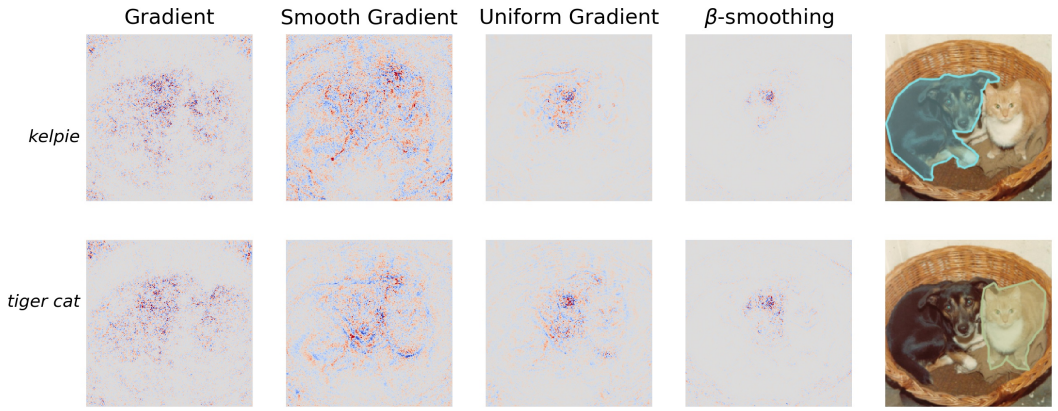

Figure 14

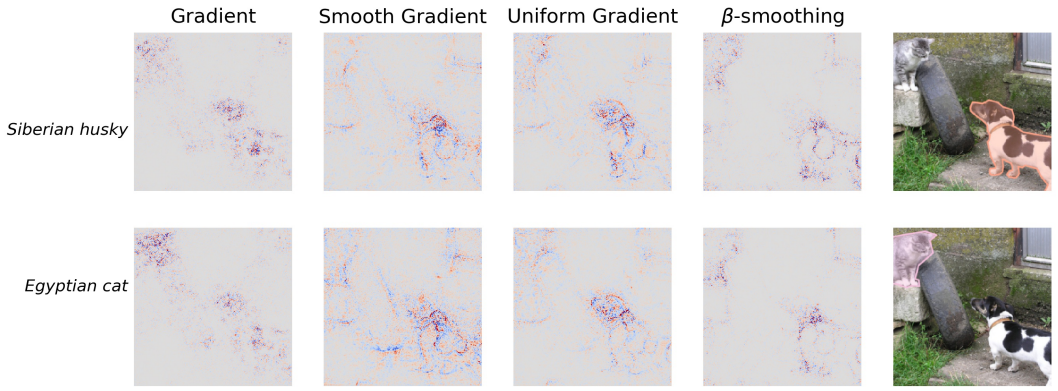

Figure 15

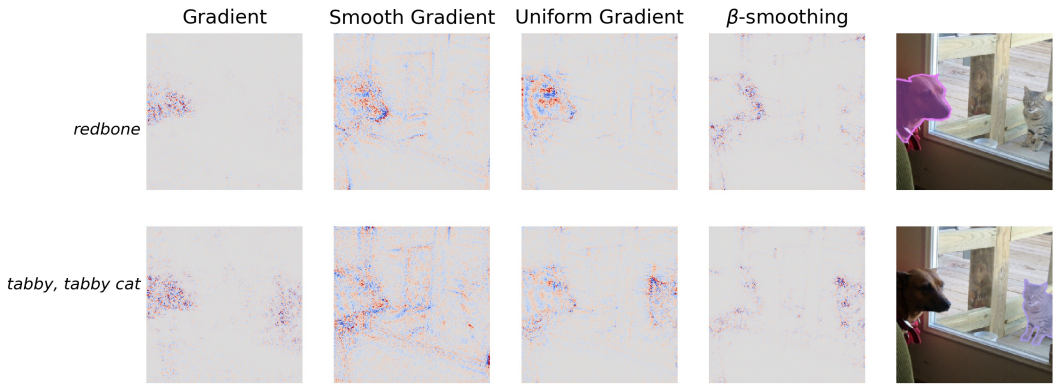

Figure 16

H ADDITIONAL EXAMPLES OF EXPLANATION ATTACKS

In this section we provide visual results for explanation attacks against post-hoc approaches for more examples.

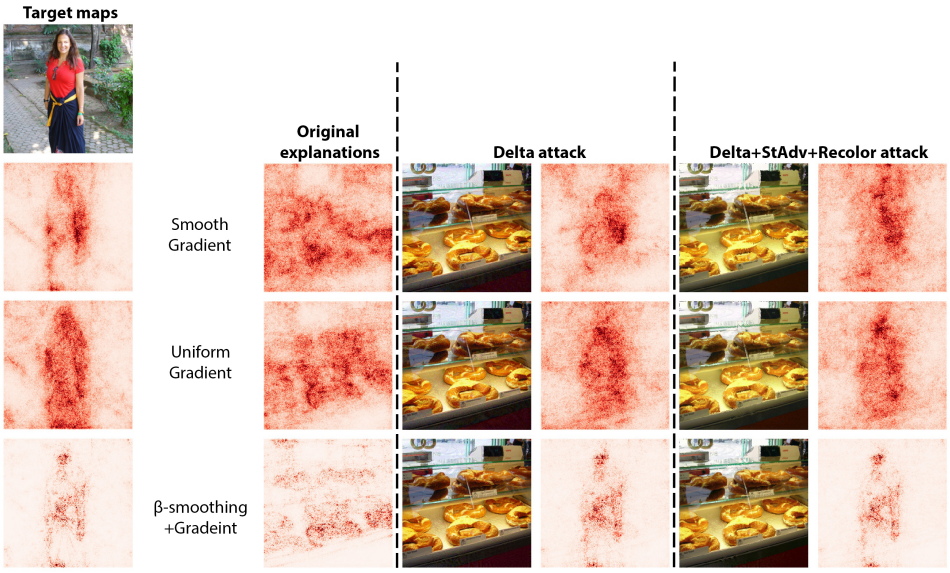

Figure 17

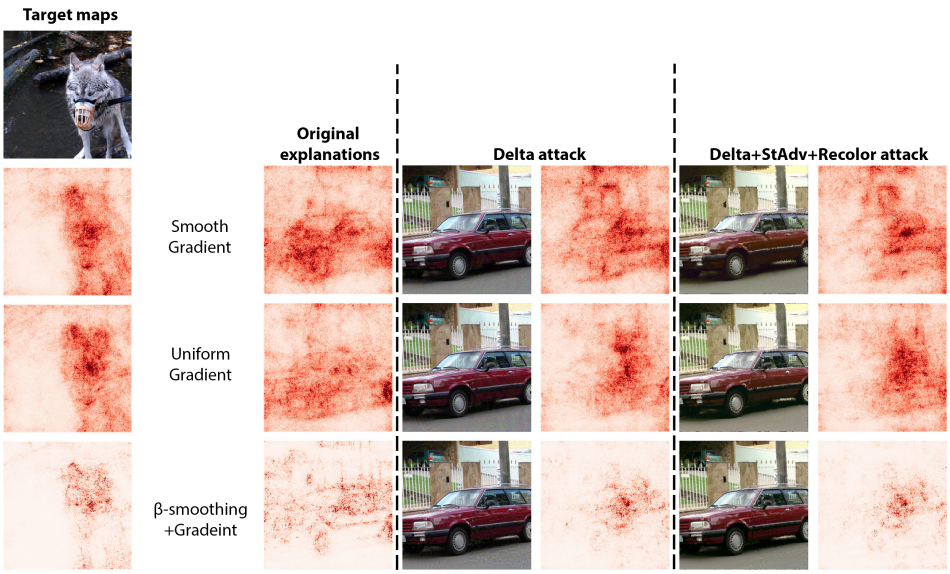

Figure 18

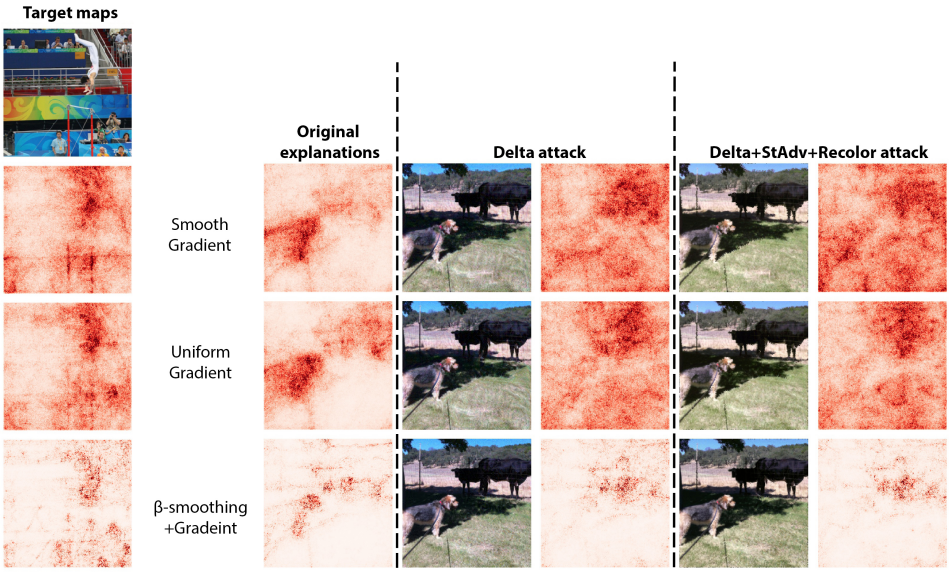

Figure 19

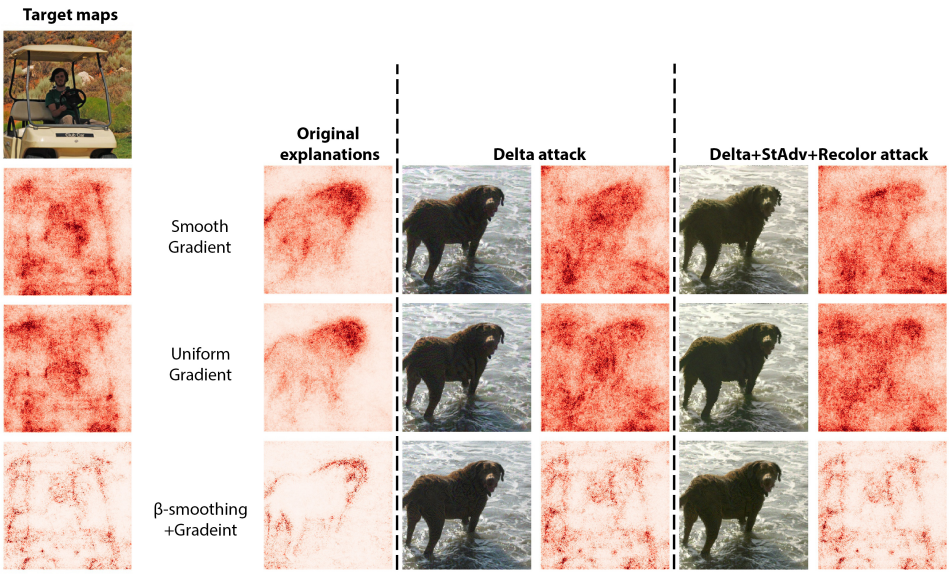

Figure 20

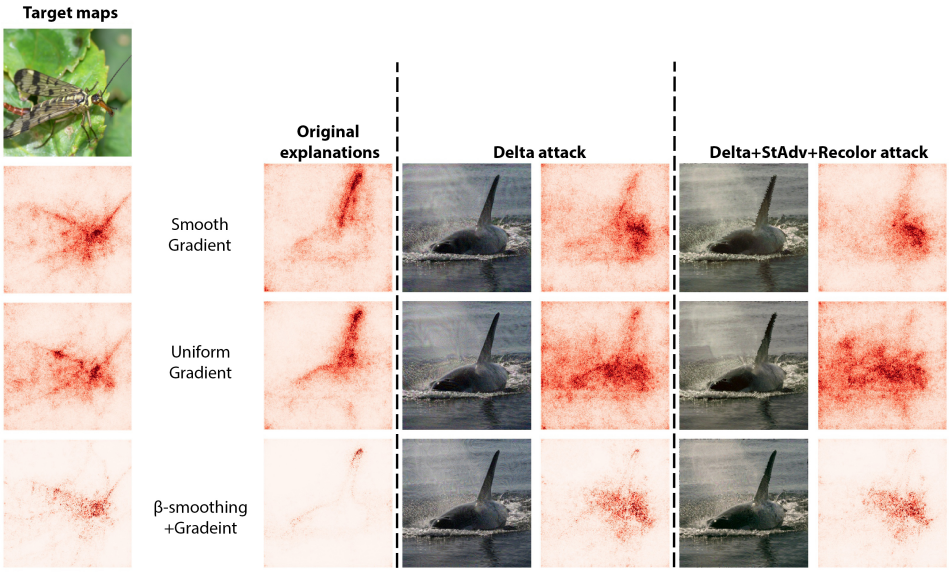
Target maps
Original explanations
Delta attack
Delta+StAdv+Recolor attack
Smooth Gradient
Uniform Gradient
β-smoothing +Gradeint

Figure 21