Saliency Methods for Explaining Adversarial Attacks

Jindong Gu
The University of Munich
Siemens AG, Corporate Technology
jindong.gu@siemens.com

Volker Tresp
The University of Munich
Siemens AG, Corporate Technology
volker.tresp@siemens.com

Abstract

In this work, we aim to explain the classifications of adversary images using saliency methods. Saliency methods explain individual classification decisions of neural networks by creating saliency maps. All saliency methods were proposed for explaining correct predictions. Recent research shows that many proposed saliency methods fail to explain the predictions. Notably, the Guided Backpropagation (GuidedBP) is essentially doing (partial) image recovery. In our work, our numerical analysis shows the saliency maps created by GuidedBP do contain class-discriminative information. We propose a simple and efficient way to enhance the created saliency maps. The proposed enhanced GuidedBP is the state-of-the-art saliency method to explain adversary classifications.

1 Introduction

The explanations produced by saliency methods reveal the relationship between inputs and outputs of the underlying model. In image classifications, the explanations are generally visualized as saliency maps. A saliency map (SM) is defined by three components: an input $\boldsymbol{x} \in \mathbb{R}^d$, a model M corresponding to a function $f_x()$, and an output class y_m.

A saliency method can be formulated as a function $g()$. A saliency map $S^m \in \mathbb{R}^d$ for the classification of the m-th class is defined as

$$S^m = g(\boldsymbol{x}, M, y_m) \tag{1}$$

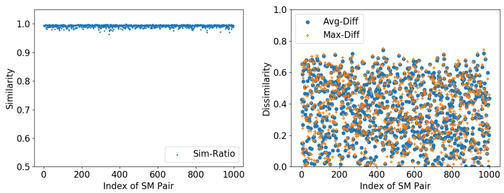
where S^m has the same dimensions as the input $\boldsymbol{x}$. The value of an element S_i^m in S^m specifies the relevance of the input feature x_i to the m-th class. To be noted that the m-th class could be neither the ground-truth class nor the predicted class y_k.

In recent years, a large number of significant saliency methods have been proposed [1–14]. [15, 16] show that SMs created by Guided Backpropagation (GuidedBP [3]) are neither class-discriminative nor sensitive to model parameters. [17] proves that Guided Backpropagation is essentially doing (partial) image recovery, which is unrelated to the network decisions. Different from their conclusions, our numerical analysis shows that the SMs created by GuidedBP do contain class-relevant decisions.

Most of the existing saliency methods only consider the SM of the ground-truth class without considering the change of the class y_m in Equation 1. [18] shows that meaningful explanations should be robust to small local perturbations of the input. However, the small perturbation can lead the misclassification of neural networks [19, 20]. We would not expect that the explanations stay unchanged in this case since the neural networks make totally different decisions. Hence, the saliency methods should be discriminative to adversary perturbation.

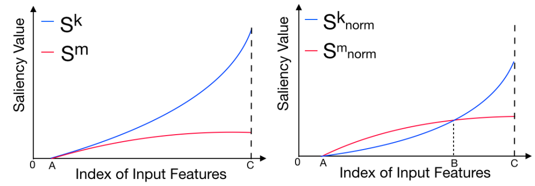
Our contributions are as follows: 1) We identify class-discriminative information in SMs created by GuidedBP and propose a simple and efficient way to enhance the created SMs; 2) We explain classifications of adversary images with the proposed enhanced Guided Backpropagation and the existing ones. Their created evaluations are evaluated via qualitative and quantitative experiments.

Preprint. Under review.



(a) Similarity of FE (b) Difference of SM Values

Figure 1: The relationship between two SMs of each SM pair: a) The similarity ratio of binarized SMs describes the similarity of Filtering Effects of them. b) The Avg-Diff and the Max-Diff between two unnormalized SMs are computed to describe the difference of their saliency values.



(a) SMs before Norm. (b) SMs after Norm.

Figure 2: This toy example illustrates how the proposed method works to enhance the discriminativity of SMs. The indexes located between A and B correspond to the input features relevant to the m-th class, and the ones between B and C are the input features relevant to the k-th class.

2 Enhanced Guided Backpropagation

Similar to raw gradient backpropagation, GuidedBP propagates gradients back to inputs and takes the received gradients as their saliency values. The two methods differ only in handling ReLU layers. In GuidedBP, $G^l = G^{l+1} * \mathbf{1}_{G^{l+1}>0 \ and \ X^l>0}$ where G^l is the gradients of the l-th layer and the X^l are the activations before RuLU layer, and $\mathbf{1}$ is the indicator function. Since the indicator function filter out part of gradients, the gradients received by some input features can be zeros, which is called filtering effect (FE). The filtering effect of a SM is formally defined as $S^m * \mathbf{1}_{S^m>0}$.

[18] provides a theoretical analysis of GuidedBP. They show that the created SMs of different classes have similar filtering effects, which means that GuidedBP is not class-discriminative. In the following, we show the SMs created by GuiedBP do contain class-discriminative information and propose a simple way to enhance the discriminative information in the corresponding saliency maps.

2.1 Identifying Discriminative Information

S^k and S^m are the two saliency maps created by GuidedBP for the k-th output class and the m-th output class. They have similar filtering effects, as theoretically analyzed in [18]. The difference between them can only be their saliency values, if existing. However, in all published work, SMs are visualized by normalizing saliency values in a SM and mapping them to a color map $[0, 255]$. The possible difference of their saliency values is hidden by the normalization.

We take pre-trained VGG16 [21] model and fine-tune it on the PASCAL VOC2012 [22] dataset. Each image in the dataset may have many objects belonging to more than one class. We select images with multiple labels from the validation dataset. For each image, we produce n SMs for n ground-truth classes and choose any two of n SMs to form a SM pair (S^k and S^m), i.e., C_n^2 SM pairs.

For each pair, we binarize the SMs and compute the similarity between two binarized ones, which is defined as the ratio of the number of pixels with the same value to the number of all pixels. All the scores of all images from the validation dataset are shown in Figure 1a. All the scores are close to 1, which means the SMs of different classes have almost the same filtering effect.

Without modifying values of SMs, we compute their average and maximum. For each SM pair, we compute the difference of saliency values of two SMs as Avg-Diff = $\frac{|Avg1-Avg2|}{\max(Avg1,Avg2)}$ and Max-Diff = $\frac{|Max1-Max1|}{\max(Max1,Max1)}$. The scores are visualized in Figure 1b. They vary from 0 to 0.8. Given a classification, the two SMs S^k and S^m differ in saliency values instead of filtering effect.

2.2 Enhancing Discriminativie Information of Saliency Maps

In this section, we propose a simple and efficient way to extract information about the difference. We argue that the relatively larger saliency values in SMs correspond to the input features that support a specific class. We extract such class-relevant information by normalize them and subtract them, which is visualized in Figure 2. Figure 2a shows the saliency values of two SMs where input features are ordered by the saliency values of a SM S^k. The two SMs have zeros in the interval [0, A] since

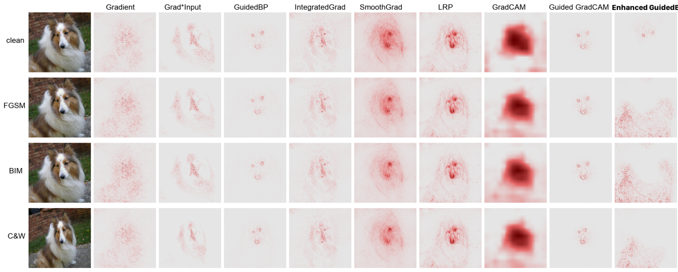


Figure 3: This figure shows SMs of clean image and adversary ones. The first column lists the original image and its adversary ones. Our Enhanced GuidedBP reacts the adversary attacks strongly, while all other the SMs produce similar SMs.

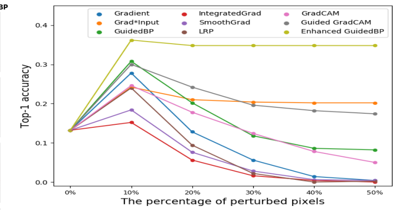


Figure 4: Following the rank of saliency values of a SM, a certain percentage of pixels of the adversary image are perturbed. The classification accuracy on the perturbed adversary images are shown.

both have the same filtering effect. The difference between the two SMs is their saliency values in the interval (A, C]. Figure 2b shows the normalized saliency values where the input features of (A, B] are relevant to the m-th class, and the ones in (B, C] are relevant to the k-th class.

In classifications of real-world images, the obtained discriminative pixels $max(0, (S^k_{norm} - S^m_{norm}))$ strongly depends on how the SMs are normalized. The trivial normalization is to divide the SM by their maximum. However, the maximal value of the SMs (i.e., the maximal local gradient value in vanilla Gradient approach) are noisy and often outliers [7, 19].

One alternative is the energy-based normalization. The individual SMs are normalized by the sum of its saliency values $|S^k|$ (i.e., the energy of the SMs). The SMs $S^k = (S^k_r, S^k_g, S^k_b)$ and $S^m = (S^m_r, S^m_g, S^m_b)$ are composed of three channels. The discriminative pixels for the k-th class on the R channel are $Dis^k_r = max(0, (\frac{S^k_r}{|S^k|} - \frac{S^m_r}{|S^m|})) = max(0, (\frac{S^k_r}{|S^k_r|+|S^k_g|+|S^k_b|} - \frac{S^m_r}{|S^m_r|+|S^m_g|+|S^m_b|}))$.

Neural networks have different sensitivity to different feature maps and input channels. In a classification, the sensitivity of channels could be different for different output classes. E.g., in case of $\frac{|S^k_r|}{|S^k_r|+|S^k_g|+|S^k_b|} \ll \frac{|S^m_r|}{|S^m_r|+|S^m_g|+|S^k_b|}$, the discriminative region $Dis^k_r = 0$, and we lose all the information on the red channel. On the contrary case, we might keep too much detail information without highlighting discriminative features. On other channels, we could similarly lose all the information or keep too much non-discriminative information.

We propose the channel-wise energy-based normalization to circumvent the problem. We consider three channels separately. The discriminative pixels of R channel is $Dis^k_r = max(0, (\frac{S^k_r}{|S^k_r|} - \frac{S^m_r}{|S^m_r|}))$. Similarly, the discriminative information of each channel is accurately identified. The generalization of the proposed enhancing method to other saliency methods will also be discussed in Section 4.

3 Explaining Classifications of Adversary Images

Inputs with imperceptible perturbation can fool the well-trained neural networks. The Fast Gradient Sign Method (FGSM) [19] perturbs an image to increase the loss of classifier on the resulting image. The Basic Iterative Method (BIM) [23] extends FGSM by taking multiple small steps instead of one big step. Another superior attack method is the Carlini and Wagner attack (C&W) [24]. In the wake of defensive distillation, they create the quasi-imperceptible perturbations by restricting their l_0, l_2 and $l_{\inf}$-norms. The l_2-norm is used across this paper.

For ImageNet validation images, we create adversary images using the three described attack methods on pre-trained VGG16. The SMs of clean images and adversary images are shown in Figure 3. For all the saliency methods except for our enhanced GuidedBP, the SMs created for predicted classes of the clean image and its adversary versions are visually the same. One might argue that it is an advantage of the saliency methods: they can still identify the object in the image even when attacked. However, we argue that saliency methods should reflect the strong reaction of deep neural networks. In other words, they should produce different SMs for clean images and adversary ones.

Since the existing saliency methods always create similar SMs for a clean image and its adversary versions, they cannot be applied to explain classifications misled by adversary perturbations. Our enhanced GuidedBP can identify the relevant evidence of the decisions. For the classification of the original input (e.g., sheepland dog), the created SM shows the VGG16 focus on the important visual feature of the target object (the head), while it focuses on class-irrelevant features (background and body parts) when explaining the classifications of adversary inputs.

The saliency methods can identify the input features that contribute to the classification decision. We can apply saliency methods on misled classifications of adversary samples. If we perturb the pixels relevant to the misclassification according to the created SMs, the attack effectiveness will be decreased. The performance of the model on the perturbated samples can be recovered to some extent. Figure 4 shows the performance of the model on the adversary samples (C&W attack) when they are perturbated according to the SMs. We can observe that the perturbation with SMs of our enhanced GuidedBP can recovery the most score. Instead of claiming the SM-based perturbation is an effective defense method, we aim to show that SMs created by enhanced GuidedBP can better identify the pixels relevant to classifications. When too many images pixels are perturbated, the visual features of true target objects is lost, which can also lead to low performance of the model.

To further analyze the adversary-discrinativity of SMs created by enhanced GuidedBP. We categorize created adversary images into two categories: Adv_s the ones that mislead the classification decisions successfully and Adv_f the ones that fail to attack the neural network. For the clean images and the perturbed images in Adv_f, the created SMs should identify the class-discriminative parts. Contrarily, for the adversary images Adv_s, the parts identified in the SMs are irrelevant to the ground-truth label, which means the network focuses on the wrong parts of the adversary images when making decisions.

Figure 5: The figure shows SMs created by GuidedBP and Enhanced GuidedBP for clean images and adversary ones. The predictions under the map indicate the success or failure of adversary attacks.

In Figure 5, the image in the first row contains a vulture. If the created adversary image fails to fool the neural network, the corresponding SM focuses on the head of the vulture (see 1st-3rd columns right of the image). If the attack is successful, the create SM for the misclassified class (i.e., kite) focuses on wings of the vulture. As a comparison, the GuidedBP always visualizes all the salient low-level features of all the images (e.g., the ski, the persons, and the alp in the image of the second row).

4 Discussion and Conclusion

Why enhanced GuidedBP is better? The pre-softmax scores (logits) are often taken as output scores to create SMs. The previous attribution methods show that the scores of different classes can be attributed to the same pixels. They explain where the scores themselves come from. Our approach explains where the difference between logits comes from, which is the exact reason why the network predicts a higher probability for a class A than a class B. In the optimization of creating adversary images, the loss of the neural network is increased, which results in the change of the rank of logits. Our approach can find the evidence for the difference between the scores, i.e., the rank of logits. The change of the rank is the reason for misclassifications. That is why the enhanced GuidedBP can explain the classification decisions of adversary images better.

The generalization of the enhancing method As analyzed in Sec. 2.1, the important factors to support the success of enhanced GuidedBP is that S^k and S^m have similar Filtering effect. When generalizing the enhancing method to other methods, the effectiveness depends on whether their created SMs have similar filtering effects.

In this work, we identify the class-discriminative information in SMs created by GuidedBP and propose a simple way to enhance it. The proposed enhanced GuidedBP can explain classification decisions of adversary images better. In future work, we will investigate how to regularize the deep neural networks using the captured discriminative information so that the rank of logits is not easily changed by adversary perturbations.

References

[1] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2013.

[2] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[3] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR*, 2014.

[4] 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):e0130140, 2015.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*, pages 1135–1144. ACM, 2016.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[7] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[8] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[10] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. A unified view of gradient-based attribution methods for deep neural networks. In *NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning*. ETH Zurich, 2017.

[11] Luisa M. Zintgraf, Taco Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *ICLR*, 2017.

[12] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NeuIPS*, pages 6967–6976, 2017.

[13] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *ICCV*, pages 3449–3457, 2017.

[14] Jindong Gu, Yinchong Yang, and Volker Tresp. Understanding individual decisions of cnns via contrastive backpropagation. In *ACCV*, 2018.

[15] Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *ECCV*, 2016.

[16] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, pages 9525–9536, 2018.

[17] Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *2018 Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

[18] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. In *Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[20] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

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

[22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[24] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.