DELVE INTO THE LAYER CHOICE OF BP-BASED ATTRIBUTION EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many issues in attribution methods have been recognized to be related to the choice of target layers, such as class insensitivity in earlier layers and low resolution in deeper layers. However, as the ground truth of the decision process is unknown, the effect of layer selection has not been well-studied. In this paper, we first employ backdoor attacks to control the decision-making process of the model and quantify the influence of layer choice on class sensitivity, fine-grained localization, and completeness. We obtain three observations: (1) We find that energy distributions of the bottom layer attribution are class-sensitive, and the class-insensitive visualizations come from the presence of a large number of class-insensitive low-score pixels. (2) The choice of target layers determines the completeness and the granularity of attributions. (3) We find that single-layer attributions cannot perform well both on the LeRF and MoRF reliability evaluations. To address these issues, we propose TIF (Threshold Interception and Fusion), a technique to combine the attribution results of all layers. Qualitative and quantitative experiments show that the proposed solution is visually sharper and more tightly constrained to the object region than other methods, addresses all issues, and outperforms mainstream methods in reliability and localization evaluations.

1 INTRODUCTION

Deep models have achieved human-level performance on a variety of computer vision tasks, while still suffering from poor decision interpretation. To address this issue, attribution methods have been widely adopted to visually identify crucial input features or regions in model decisions. Existing attribution methods can be roughly classified into two categories. One is backpropagation(BP)-based attribution, which only requires one or a few backpropagations to obtain the attribution results. This easy and lightweight implementation has made it widely used in different areas. The other is non-backpropagation attribution, such as perturbation-based interpretation (Fong & Vedaldi, 2017), Shapley value (Lundberg & Lee, 2017), information bottleneck (Schulz et al., 2019), etc. These algorithms use frequent queries or complex optimizations to obtain attributions and are high computational complexities.

In this paper, we focus on BP-based attribution methods and study the influence of the choice of the target layer. Since backpropagation can use and output the results of any layer, layer choice is naturally one of the fundamental factors of BP-based attribution methods and needs to be studied in-depth. However, few existing methods focus on the target layer. Specifically, the mainstream BP-based methods can be divided into two types (see Figure 1 and 2): (1) Gradient-based methods, which distribute the final decisions to the inputs and are characterized by noise and significant edges. (2) CAM-based methods that use the feature maps closest to the output and are characterized by smooth and complete results.

Recently, several researchers have noticed a connection between the choice of the target layer and the reliability of interpretation results. For example, Rebuffi et al. (2020) point out that the layers close to the input are class-insensitive, and Jalwana et al. (2021) find that layers close to the output have low resolution and cannot give fine-grained attributions. However, they do not explicitly investigate the nature of layer choice, but address these issues in a straightforward way, such as subtracting meta-saliency to improve class sensitivity or increasing input size to obtain finer-grained results. As

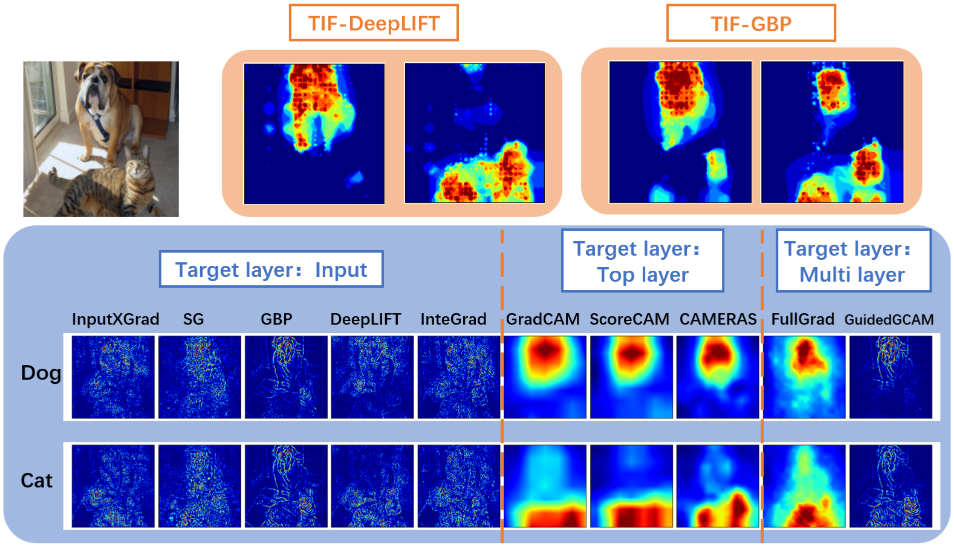


Figure 1: Intuitive comparison of TIF(α=0.7) and baseline for class sensitivity. TIF can improve the class sensitivity even if on the GBP, which are considered invariant to higher layer parameters(Adebayo et al., 2018).

a result, the study and use of the influence of the target layer are far from adequate, and it is still unclear to the user how the choice of the target layer affects the final attribution result.

Our goal is thus to give a holistic understanding of the influence of layer choice on BP-based attributions. In this work, we focus on three issues:

- Is the problem of class insensitivity and low resolution truly caused by the different choice of target layer?
- What other problems resulted from the choice of target layer?
- How should we select and improve the existing methods from the perspective of the target layer?

To answer these questions, we make the following contributions in this work:

Quantifications by the backdoor attack. Inspired by evaluation through backdoors(Lin et al., 2020), we propose that backdoor attacks can be used to control the experimental variables and allow the model to have known and determined decision behaviors. As in Figure 3, we design three different backdoor triggers and quantitative metrics for the three target properties: class sensitivity, fine-grained localization, and completeness.

Studying the impact of layer choice. Using the quantification of middle-layer attributions of gradient-based methods, we show that energy distributions of the bottom-layer attribution are class-sensitive, and the class insensitive visualizations come from the presence of a large number of class-insensitive low-score pixels. Moreover, the choice of target layer actually affects the fine-grained localization capability and the completeness of the target objects: the bottom layer enables fine-grained localization but poor completeness, while the top layer is the opposite (see Figure 4). We further find that this property leads to trade-offs in different reliability evaluations for single target layer attribution: the bottom attributions perform well on MORF(Samek et al., 2016) while the top attributions are good at LeRF (see Figure 5 and 6).

Fusing different layers to obtain better explanations. Based on the properties of attributions of the different target layers, we propose a new strategy for fusing all layers, TIF (Threshold Interception and Fusion), by intercepting and fusing the certain threshold values of the attributions of each layer. Qualitative and quantitative experiments show that the proposed solution is visually sharper and more tightly constrained to object regions than other methods (Figure 1), addressing issues

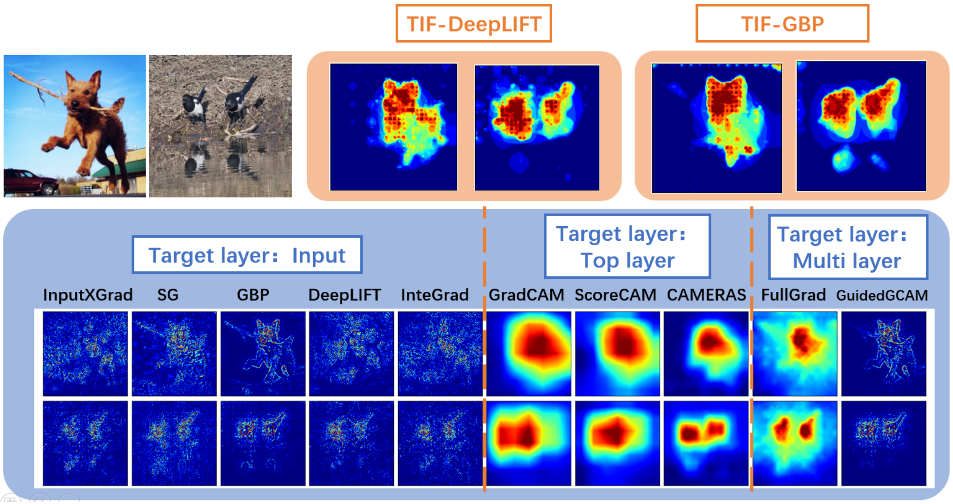


Figure 2: Intuitive comparison of TIF(α=0.7) and baseline for completeness and fine-grained localization. Since layer choice determines the completeness and granularity of attribution, our layer fusion strategy consistently improves completeness and obtains fine-grained results for different BP-rules.

from layer choice (Figure 2) and outperforming mainstream methods in reliability (Figure 5) and localization evaluation (Table 2).

2 Related Work

The goal of BP-based attribution is to capture the importance of inputs for decision-making by modifying the backpropagation rules of the model. Depending on the choice of the target layer, there are three types of BP-based attribution methods: gradient-based, CAM-based, and combined.

Gradient-based attributions. Gradient-based attributions practically distribute the final decision to the input pixels. The absolute value of the original gradient (Simonyan et al., 2013) is the earliest saliency map, which is intuitive but overly noisy. Deconv (Zeiler & Fergus, 2014), GBP Springenberg et al. (2014), shields negative gradients to improve visualizations. DeepLIFT Shrikumar et al. (2017) and IntegratedGradient (Sundararajan et al., 2017) notice the gradient saturation problem and introduce reference point and path integral solutions. These methods select the input layer as the target layer.

CAM-based attributions. CAM-based, also known as activation-based attributions, aims to combine weighted activation maps and usually uses the last convolutional layer before the global pooling layer. GradCAM Selvaraju et al. (2017) use gradient to weight the top layer activations and proposes that the convolutional layer has a strong spatial prior and can simply use bilinear upsampling to obtain the same size attributions as the inputs. Score-CAM(Wang et al., 2020) propose the increase of confidence rather than gradients. CAMERAS Jalwana et al. (2021) claims that GradCAM is low resolution due to downsampling, so it increases the resolution of the input. These methods select the top layer as the target layer.

Combinations of different layers. some researchers recognize the limitation of one target layer. Guided GradCAM(Selvaraju et al., 2017), use GradCAM×GBP to obtain both the class sensitivity of GradCAM and the fine-grained localization of GBP. FullGrad(Srinivas & Fleuret, 2019) concern that the attributions of biases also need to be considered, and therefore fuse the attribution results of different layers of bias. However, in Figure 1, 2, and 5, we show that these methods do not effectively address the problems caused by single target layer attributions.

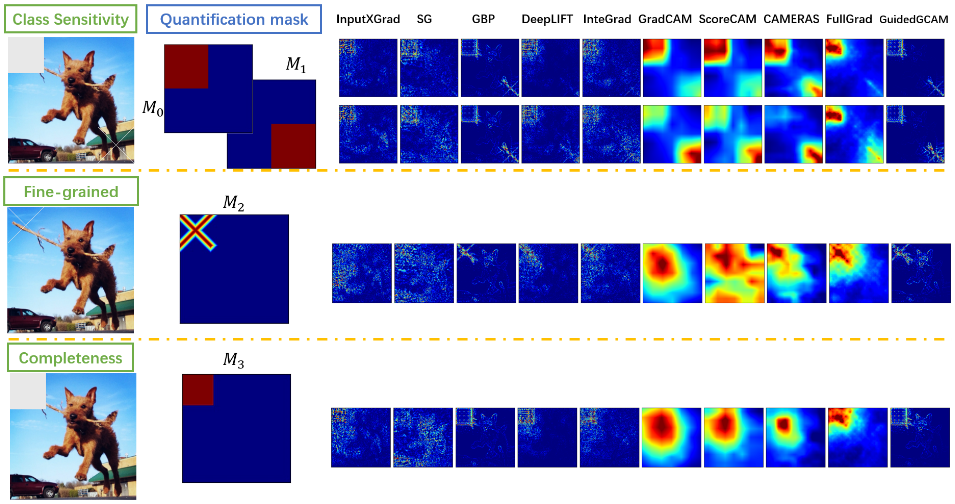


Figure 3: Triggers, quantification masks, and intuitive results on baseline for the backdoor attacks on class sensitivity, fine-grained localization, and completeness.

3 Class Sensitivity and Fine-grained Localization

3.1 Quantification by Backdoor Attacks

Class insensitivity and low resolution are frequent problems in attribution methods and can gravely affect reliability. Past researchers have recognized that these issues may be related to the choice of the target layer, but due to the lack of decision ground truth, it is difficult to clarify and quantify these issues in practice.

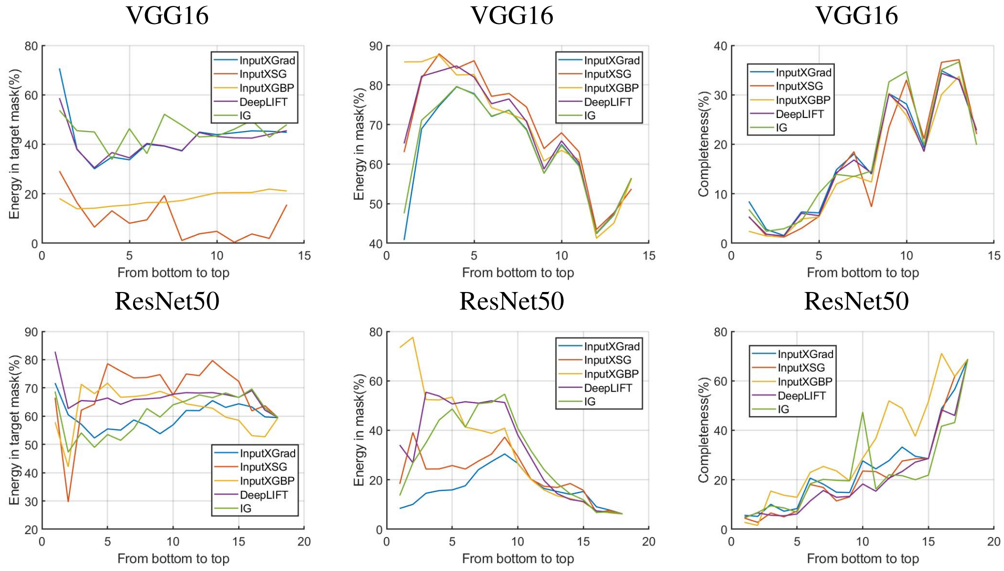
To address this issue, we use backdoor attacks (Liu et al., 2020) to control the decision-making process (See Figure 3). Backdoor attacks refer to adding predefined backdoor triggers to the training set samples, such that the final trained model will predict a certain label when encountering any input with such a trigger. This property allows us to obtain a reliable ground truth for the model decision: the model decision must be derived from the backdoor trigger. From this, we can easily design trigger-based metrics to evaluate the performance of attributions. Moreover, by looking inside the changes in the metrics used for intermediate layer attributions, we can easily determine if the choice of target layer is a key factor in these problems. In the subsequent description of the quantification process of the backdoor attack, we are committed to making clear two key points: (1) What is the trigger used? (2) How to design metrics?

Setup. See Appendix A.1 for backdoor experimental details. We provide more results for VGG16 (Simonyan & Zisserman, 2014) and ResNet50 (He et al., 2016) on ImageNet in Appendix B.

3.2 Class Sensitivity

Motivation. The class insensitivity problem is that the same input with different decisions corresponds to very similar attributions and does not correspond to the target objects. For example, in Figure 1, there are a dog and a cat, but gradient-based methods, which use the input as the target layer, do not highlight the dog or the cat according to the label.

Quantification. To verify this problem, we add two triggers in the model: a white square to the top left corner and a white cross to the bottom right corner. During training, square triggers are added with label 0 and cross triggers are added with label 1. For evaluation, we add both triggers in the image and obtain the attributions with label 0 and label 1. If an explanation is class-sensitive, it must highlight the top left corner of label 0 and the bottom right corner of label 1. Therefore, we design a metric to show how much energy of the attribution map around the target corner to represent the class sensitivity :$\frac{1}{2}(\frac{\sum(M_0*R_0)}{\sum R_0}+\frac{\sum(M_1*R_1)}{\sum R_1})$, where R_0 and R_1 is the attributions for label 0 and



(a) Class sensitivity (b) Fine-grained localization (c) Attribution completeness

Figure 4: The quantitative evaluation of middle-layer attributions for class sensitivity, fine-grained localization, and attribution completeness. The latter two of them exhibit significant correlations with layer choice.

label 1, M_0 and M_1 is the mask provided in Figure 3, the implementation of M_i can be found in Appendix A.1.

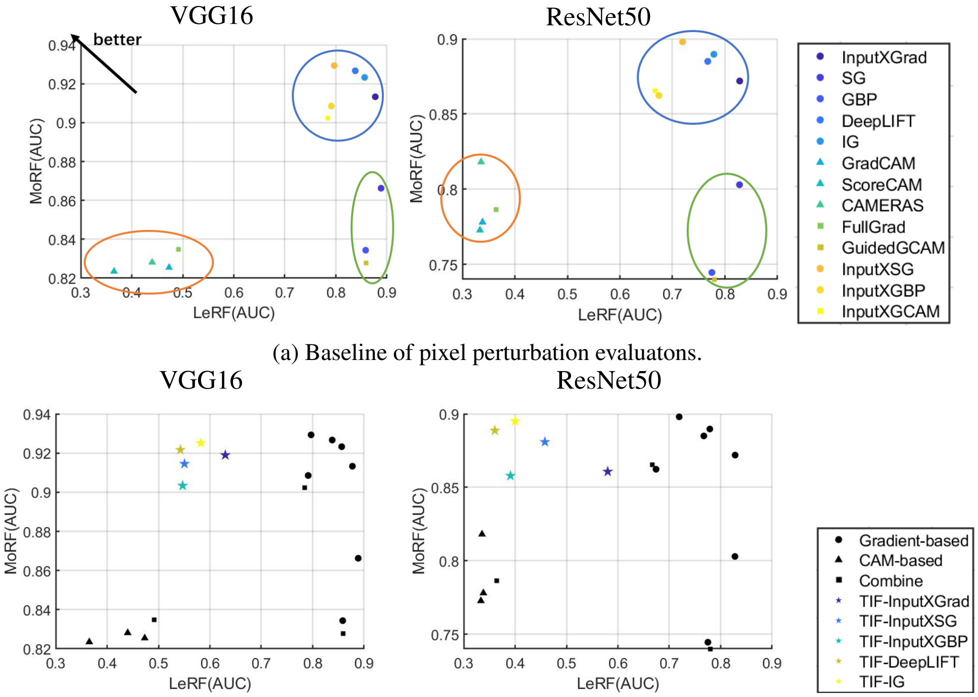
Middle layer analysis. We obtain the intermediate layer results by directly truncating the model into two halves from the target layer and then running gradient-based attribution results for the top half. Such a design has the advantage of not requiring any modifications to the internal implementation of the algorithm and guarantees consistency. Since the intermediate layers normally have more than 1 channel and smaller sizes, we simply sum and bilinear sample to obtain the attribution score for each input pixel. The formalization of this operation can be seen in Section 5. Note that we multiply the activations for SmoothGrad and GBP (denoted as Input×SG and Input×GBP) because their middle layer attributions are all nonpositive. Considering that all other methods are designed to naturally multiple activations and provide reasonable results, we modify them to be consistent.

Results. The quantitative results are shown in Table 1 and Figure 4(a). It can be seen that the class sensitivity does not have a significant trend response to changes in the target layer, meaning that it is not determined by the choice of the target layer, but by the choice of the algorithm. The previous misunderstanding of the underlying class insensitivity most likely stems from the fact that a small number of class-relevant strong attributions are distributed over discrete pixels and would be ignored by humans as noise, while a large number of class-independent weaker attributions appear to be particularly prominent. See Appendix B.1 for more intuitive results on the baseline and intermediate layer attributions.

3.3 Fine-grained Localization

Motivation. As downsampling is widely adopted in convolutional neural networks, the activation maps of the top layer are low-resolution, so attribution results do not provide a sufficiently fine-grained result. Moreover, the top layer activations may not correspond exactly to the input pixels, and these issues can severely hinder CAM-based methods from providing fine-grained localization. The intuitive results in Figure 3 show this failure of localization of CAM-based attributions, especially face multi-target objects.

Quantification. To quantify this problem, we inject a white cross trigger on the top left corner of the model (see Figure 3), such trigger is only 1 pixel width to ensure that the decision basis of the model is sufficiently fine-grained. The corresponding mask M_2 is also located at the top left corner, then



(a) Baseline of pixel perturbation evaluatons.

(b) Comparison of attributions before and after TIF.

Figure 5: Pixel perturbation-LeRF/MoRF, The closer the method is to the top-left the better.

we can use the energy attributed to the mask to evaluate the performance of fine-grained localization: $\frac{\sum(M_2*R)}{\sum R}$, where R is the attribution results.

Results. The settings of middle-layer analysis are the same as Section 3.2, and we can obtain the quantitative results in Table 1 and Figure 4(b). Apart from the results for a few bottom layers, the experimental results as a whole demonstrate a consistent and significant negative trend in correlations, suggesting that the choice of target layer does affect the fine-grained localization performance and that top-layer attributions suffer from low attribution accuracy. See Appendix B.2 for more intuitive results on the baseline and intermediate layer attributions.

4 Completeness and Reliability

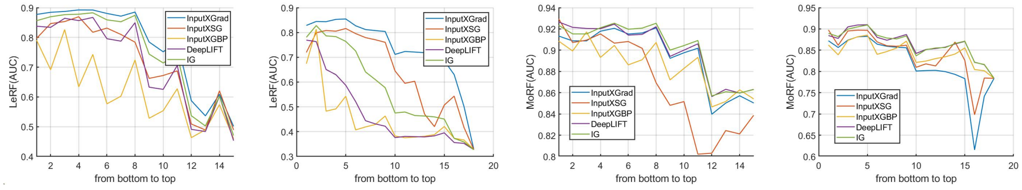
What further problems arise from the choice of target layer? In this section, we find that the completeness and reliability of attributions strongly depend on the choice of the target layer, and that one-layer attributions do not provide a satisfactory explanation.

4.1 Completeness of Attributions

Motivation. Intuitively, gradient-based methods tend to provide noise and edges, rather than complete objects. In contrast, CAM-based attribution methods consistently provide complete regions. Therefore, completeness is likely to be a property that depends on the target layer and deserves quantitative verification.

Quantification. To study the completeness, we propose a simple white square trigger (Figure 3). We concentrate on whether the attributions can provide a complete square trigger and the quantification mask M is just the top left square. The metric is just the average energy in the mask: $\frac{\sum(M_3*R)}{\sum M_3*max(M_3*R)}$.

Results. As shown in Table 1 and Figure 4(c), completeness is heavily influenced by the target layer and the higher layers have consistently better performance. Since many previous studies have



(a) LeRF+VGG16 (b) LeRF+ResNet50 (c) MoRF+VGG16 (d) MoRF+ResNet50

Figure 6: The quantitative evaluation of middle-layer attributions for LeRF and MoRF.

Model	VGG16			ResNet50		
Quantifications(↑)	Class	Fine	Comple	Class	Fine	Comple
InputXGrad	0.70	0.42	0.085	0.71	0.12	0.06
SG	0.23	0.56	0.054	0.52	0.22	0.05
GBP	0.18	**0.83**	0.024	0.54	**0.71**	0.03
DeepLIFT	0.57	0.66	0.055	0.81	0.39	0.05
IG	0.52	0.48	0.071	0.70	0.23	0.04
GradCAM	0.41	0.36	0.48	0.51	0.06	0.77
ScoreCAM	**0.84**	0.26	0.80	0.58	0.06	0.69
CAMERAS	0.65	0.38	0.61	0.54	0.08	0.41
FullGrad	0.53	0.20	0.64	0.47	0.07	0.72
GuidedGCAM	0.23	0.57	0.17	0.77	0.35	0.15
TIF-DeepLIFT(0.5)	0.45	0.12	**0.83**	0.57	0.12	**0.81**
TIF-DeepLIFT(0.7)	0.54	0.2	0.78	0.68	0.17	0.66
TIF-DeepLIFT(0.9)	0.48	0.48	0.52	**0.83**	0.29	0.28

Table 1: Quantification evaluations of class-sensitivity(Class), fine-grained localization(Fine), and completeness(Comple). TIF achieve competitive performance especially in completeness.

mentioned that the bottom layer prefers edges and textures, while higher layer features usually correspond to higher level semantics, the lack of completeness in the bottom layer is natural. In fact, from the intuitive results of the hierarchical analysis, we can also see how the model gradually processes the various edges of the input to complete the square trigger. See Appendix B.3 for more intuitive results on the baseline and intermediate layer attributions.

4.2 Reliability of Attributions

The reliability of attribution methods has long been of interest, and the most popular reliability evaluation metric is pixel perturbation (Samek et al., 2016), where the importance of these pixels in the decision can be judged by replacing some pixels in the input with uninformative value (usually 0) and observing the change in the model output. There are two implementation strategies for common pixel perturbations, one is LeRF, where the least salient pixels are removed first and reliable attributions ensure those model decisions do not change. The other is MoRF, where the most salient pixels are removed first so that reliable attributions can change the decision by removing a very small number of pixels. Therefore, when we draw the curve with the horizontal coordinate as the pixel removal rate and the vertical coordinate as the decision change rate, the area under the curve (AUC) can be used to quantify the performance, and lower LeRF(AUC) is better while higher MoRF(AUC) is better. See Appendix A.2 for experimental details.

Figure 5(a) shows the results of the evaluation metrics for each baseline method. Note that the closer the top left corner, the better the method. It can be seen that the performance of the methods is significantly divided into three parts. The blue circles are the methods with good MoRF results and poor LeRF results, which are mostly gradient-based methods. The orange circles with good LeRF results and poor MoRF results are mainly CAM-based methods. However, the green circle has both poor LeRF and poor MoRF results. We find that the methods in the green circles, namely SG, GBP, and GuidedGradCAM, do not multiply the input, so they cannot represent the change of setting pixels to zero. After multiplying the inputs, all three methods have better MoRF and run

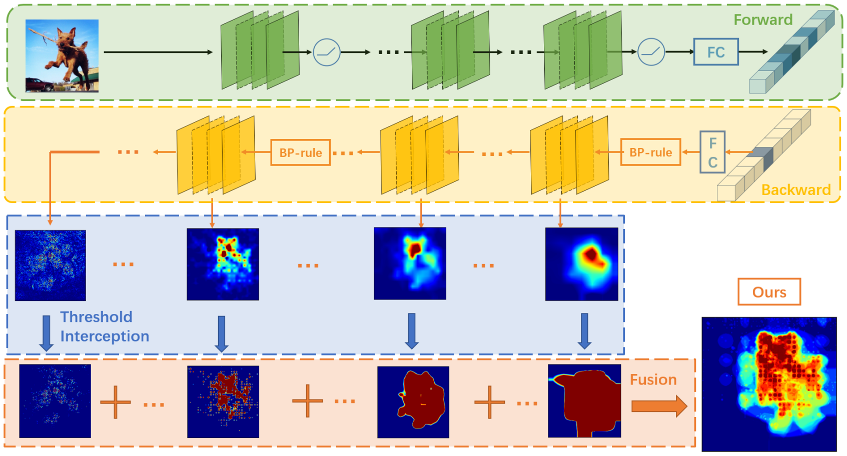


Figure 7: Pipeline of Threshold Intercetion and Fusion(TIF).

into the blue circle. This phenomenon explains to some extent why we need to use Input×SG and Input×GBP in middle layer analysis.

The significant difference in MoRF and LeRF between Gradient-based and CAM-based methods suggests that reliability also depends on the choice of the target layer. Then we evaluate middle-layer attributions and propose the results in Figure 6. It can be seen that the higher the target layer, the better the LeRF, but the worse the MoRF. This phenomenon is most likely related to fine-grained localization and completeness, which have similar trends as LeRF or MoRF. For LeRF, if the attribution is incomplete, the least salient removal may corrupt the target object by removing pixels from it. As for MoRF, the weakness in fine-grained localization makes it difficult to accurately correspond to the few most important key points.

5 Threshold Intercetion and Fusion

In Section 3 and Section 4, we demonstrate the impact of layer choice: higher layers have better LeRF and completeness, but worse MoRF and fine-grained localization performance. Moreover, the bottom attributions are full of class-insensitive low-score pixels, while the top attributions cannot accurately locate the key pixels. Therefore, we need to fuse the layers to obtain the overall performance improvement. We propose Threshold Interception and Fusion(TIF), which aims to summarize the high-score region of the bottom layer attribute and the low-score region of the top layer attribution. The overall pipeline is shown in Figure 7, and next, we will describe the details of the method implementation.

Target layer attribution. Let x be the input image and neural network f have L layers, A_l is the activation of l^{th} layer. Denote $r_k^{(l)} = \frac{\partial f_k}{\partial A_l}$ as the attribution results of l^{th} layer of label k, and $r_k^0 = r_k$ is the attribution of inputs. $r_k^{(l)}$ have the same shape(BCHW) of activations A_l, i.e. the same batch sizes, channels, heights, weights.

Channel reduce. Note that the channel C of the middle layer is large, and we only need saliency maps of the same size as the input, we sum up the channels C of r_k^l to obtain middle-layer attributions which can provide an important score for each location $R_k^{(l)} = \sum_{c \in C} r_k^{(l)}$.

Up-sampling. The middle layer activations A_l usually have smaller heights and weights than input x, so we need to upsampling the attributions $R_k^{(l)}$ to obtain the same size as x. GradCAM uses bilinear upsampling algorithm (denoted as ϕ), and uses ReLU function σ to highlight the positive contributions, most of CAM-based methods followed such setup: $\tilde{R}_k^{(l)} = \sigma(\phi(R_k^{(l)}))$.

Threshold interception with parameters. Now we have L middle layer attributions $\tilde{R}_k^{(l)}, l = 0, 1, .., L-1$. Our algorithm aims to intercept high-score regions of the attributions close to the input and low-score regions from the top layers. After ranking the attribution scores for all pixels from high to low, we denote $s(\tilde{R}_k^{(l)}, t) = s(t)$ as the top t score value of $\tilde{R}_k^{(l)}$. For the l^{th} layer, we remove the attributions less than $s((l+1)/L)$, and clamp the scores greater than $s(l/L)$:

$$\hat{R}_k^{(l)} = minimum\left(\tilde{R}_k^{(l)}\mathbb{I}\left(\tilde{R}_k^{(l)} > s\left(\frac{l+1}{L}\right)\right), s\left(\frac{l}{L}\right)\right) \tag{1}$$

Note that $s\left(\frac{l}{L}\right)$ might be zero and corresponding layer contribute little, we use α to avoid this condition.

$$\hat{R}_k^{(l)} = minimum\left(\tilde{R}_k^{(l)}\mathbb{I}\left(\tilde{R}_k^{(l)} > s\left(\frac{l+1}{L}(1-\alpha)\right)\right), s\left(\frac{l}{L}(1-\alpha)\right)\right) \tag{2}$$

Fusion. We use the same fusion strategies as the post-process of FullGrad: z-score standardized the middle layer attributions to $[0, 1]$ and sum it up: $TIF = \sum_{l=0}^{L-1} \psi(\hat{R}_k^{(l)})$ where $\psi(t) = \frac{t-min(t)}{max(t)-min(t)}$ is z-score standardization. We then use TIF to improve the attribution methods.

6 Experiments of TIF

Next, we will experimentally verify whether TIF can improve the performance of the properties listed above. We show the results for TIF-DeepLIFT, but TIF can be applied to any Gradient-based method and achieve consistent improvements. See Appendix C for more details and experimental results.

Class sensitivity. The intuitive results of Figure 1 and quantification results of Table 1 show excellent class sensitivity of TIF-DeepLIFT on ResNet.

Fine-grained localization and attribution completeness. Figure 2 shows the intuitive performance of Fine-grained localization and attribution completeness. TIF-DeepLIFT($\alpha = 0.7$) provides complete and fine-grained attributions to the objects, e.g. the branch in the mouth of the dog is removed precisely and the two birds are clearly separated. Quantitative results demonstrate the best completeness and better fine-grained localization performance than CAM-based attributions. See Appendix C.2 and C.3 for more intuitive results.

EBPG(↑)	VGG16	ResNet50
InputXGrad	39.29	35.54
SG	44.27	45.56
GBP	51.08	52.18
DeepLIFT	45.40	36.84
IG	43.04	39.32
GradCAM	41.36	41.04
ScoreCAM	53.36	44.82
CAMERAS	40.96	48.72
FullGrad	37.76	35.78
GuidedGCAM	59.38	60.80
TIF-DeepLIFT(0.5)	40.64	53.20
TIF-DeepLIFT(0.7)	48.36	62.34
TIF-DeepLIFT(0.9)	**60.44**	**72.76**

Table 2: Comparative evaluation on Energy-Based Pointing Game.

Evaluations of reliability. The results are presented in Figure 5(b), and it can be seen that TIF consistently improves the combined performance of LeRF(AUC) and MoRF(AUC) of all methods and TIF-DeepLIFT achieves the best results. See Appendix C.4 for the analysis of hyperparameter α.

Localization evaluation. We use the energy-based point games(EBPG)(Wang et al., 2020) to evaluate the localization, which compares the proportion of energy within the target bounding box to all attributions. See Appendix A.3 for experimental detials. Table 2 shows the results on VGG16 and ResNet50. It can be seen that TIF-DeepLIFT(α=0.9) significantly outperforms all mainstream methods.

7 Conclusion

In this paper, we investigate how the choice of layers affects the attribution results. With the quantification of backdoor attack, we observe the performance of different layer attributions on class sensitivity, granularity, and completeness. In addition, we point out that single-layer attributions are not reliable on both LeRF and MoRF evaluations, and propose a novel approach to fuse all layer attributions to address the issues arising from layer choice.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

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

Mohammad AAK Jalwana, Naveed Akhtar, Mohammed Bennamoun, and Ajmal Mian. Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16327–16336, 2021.

Yi-Shan Lin, Wen-Chuan Lee, and Z Berkay Celik. What do you see? evaluation of explainable artificial intelligence (xai) interpretability through neural backdoors. *arXiv preprint arXiv:2009.10639*, 2020.

Yuntao Liu, Ankit Mondal, Abhishek Chakraborty, Michael Zuzak, Nina Jacobsen, Daniel Xing, and Ankur Srivastava. A survey on neural trojans. In *2020 21st International Symposium on Quality Electronic Design (ISQED)*, pp. 33–39. IEEE, 2020.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8839–8848, 2020.

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

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

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

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

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *arXiv preprint arXiv:1905.00780*, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 24–25, 2020.

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

A Experimental Details

A.1 Backdoor Attacks

Backdoor attacks, also known as poisoning attacksLiu et al. (2020), aims to train a model which can provide correct predictions for clean inputs, and output a given label when adding a certain trigger. Formally, attackers usually add a certain percentage of trojan samples x' to the training set. The only difference between these trojan samples x' and corresponding normal inputs x is that their label y' is fixed and have a particular trigger $x' = x \odot (1 - M_t) + \Delta \odot M_t$, where $\Delta = 1$ is the trigger and M_t is a 2-D matrix, representing a mask with values in $\{0, 1\}$. For briefly, we use $M = [condition(i, j)]$ to denote a mask $M = \{m_{i,j}\}$ that $m_{i,j} = 1$ for $condition(i, j)$ and $m_{i,j} = 0$ for others. In this work, we inject the backdoors by finetuning the pre-trained model VGG16 and ResNet50 with a step size of 1e-4, 10% data are trojans, and train 15 epochs on ImageNetRussakovsky et al. (2015) so that it achieves 100% attack success rate.

Trigger for fine-grained localization. we add X-shaped triggers to the top left corner and inject them into the VGG16 and ResNet50: $M_t = [(0 \le i, j < 64) \& (j = i \ or \ j = 64 - i)]$. The quantification mask M_2 is similar to the trigger but 16 pixel width.

Trigger for completeness. we propose a trigger which is a white square in the top left corner: $M_t = [0 \le i, j < 64]$. The quantification mask M_3 is just the trigger.

Trigger for class sensititivity. we add two triggers mentioned above to the top left and bottom right corners, and assign label 0 and label 1 to them respectively. The quantification masks M_0 and M_1 are square with 96 pixels width on the top left and square on the bottom right corners, such design can capture more energy around the trigger and mitigates the influence of lacking fine-grained details.

A.2 Pixel Perturbation Evaluations

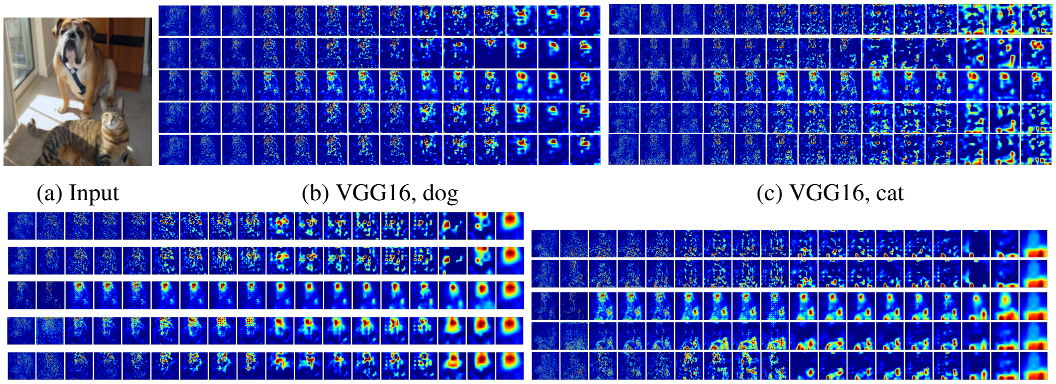
We perform two kinds of pixel perturbations: removing most relevant input features (MoRF) and removing least relevant input features (LeRF)Samek et al. (2016). Specifically, our procedure is as follows: for a given value of k, we replace the k pixels corresponding to k most/least salient values with zero pixels. We obtain the results on 50000 images of ImageNet validation set.

A.3 Energy-based Point Games

Wang et al. (2020) provide an Energy-Based Pointing Game (EBPG) to evaluate the localization ability of attribution methods. Such evaluation reflects whether the highlighting regions of method are consistent with humans. Specifically, the evaluation is attributed inside the bounding boxes as a proportion of all attributions:

$$EBPG = \frac{\sum R_{x \in bbox}}{\sum R_{x \in bbox} + \sum R_{x \notin bbox}} \tag{3}$$

We use the same setting as Wang et al. (2020): Removing images where an object occupies more than 50% of the whole image to guarantee the bounding box makes sense. We experiment on 1000 random selected images from the ILSVRC2012 validation set.


(a) Input (b) VGG16, dog (c) VGG16, cat

(d) ResNet50, dog (e) ResNet50, cat

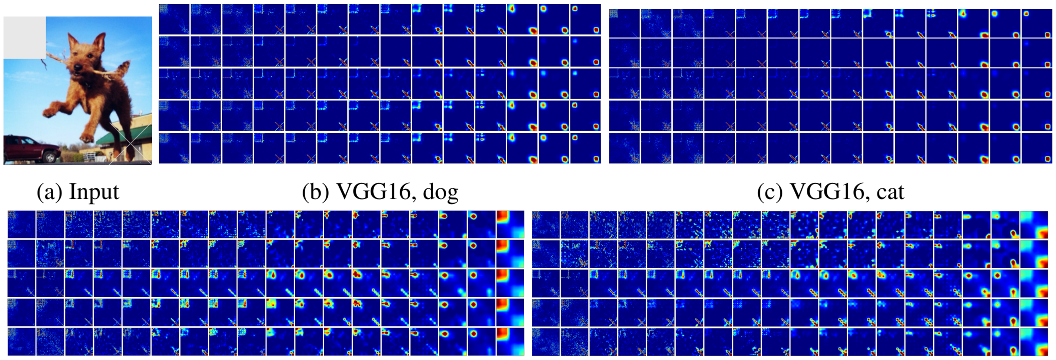
Figure 8: Middle layer attributions for InputXGrad, InputXSG, InputXGBP, DeepLIFT, Integrated-Gradient.

B Intuitive Results of Backdoor Attack

B.1 Class Sensitivity

Figure 8 show the intuitive results of class sensitivity in the middle layer. It can be seen that the bottom result is not necessarily more class insensitive than the top layer attributions, and the differences in class sensitivity between BP-rule are much larger than the differences in sensitivity due to layer choice. Except for the top layer of ResNet, because from this layer to the output is just a simple fully-connected layer with no difference between BP-rule.

Figure 9 provide the middle layer attributions of backdoor attack, which have the similar results.

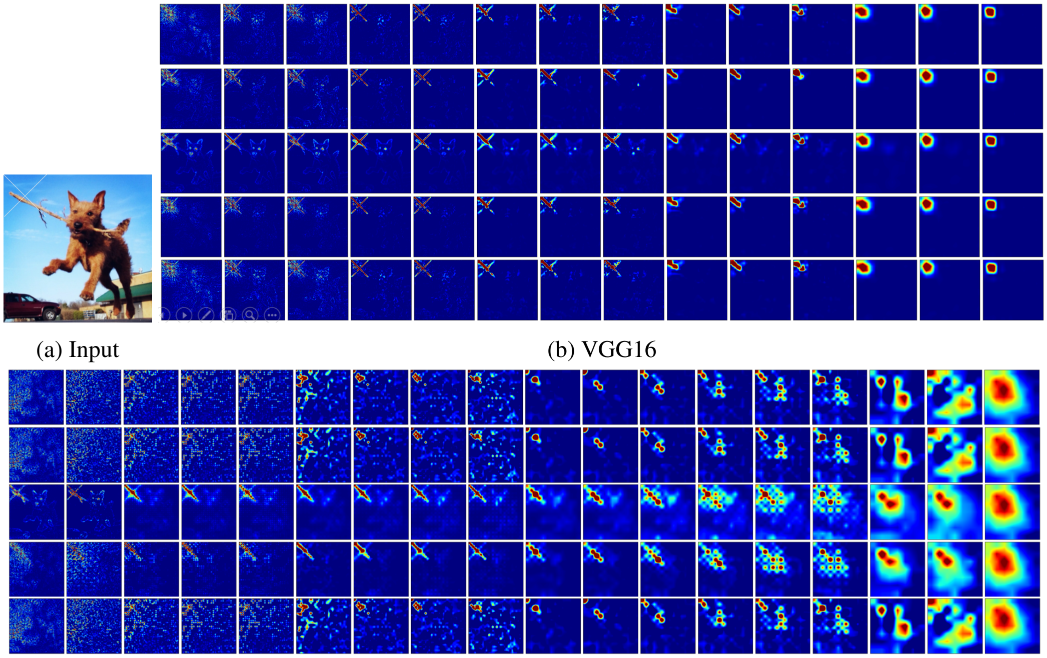

(a) Input (b) VGG16, dog (c) VGG16, cat

(d) ResNet50, dog (e) ResNet50, cat

Figure 9: Middle layer attributions for InputXGrad, InputXSG, InputXGBP, DeepLIFT, Integrated-Gradient with backdoor triggers of class sensitivity.

B.2 Fine-grained Localization

Figure 10 show the intuitive results of middle layer attributions for cross trigger. As can be seen, the bottom attribution can clearly locate the X trigger. As the layer choice rises, all kinds of methods gradually deviate from the original trigger position, especially ResNet50 which has a short cut.



(a) Input

(b) VGG16

(c) ResNet50

Figure 10: Middle layer attributions for InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient with backdoor triggers of fine-grained localization.

B.3 Completeness

Figure 11 show the intuitive results of middle layer attributions for square trigger. As can be seen, the bottom attribution can clearly locate the X trigger. As the target layer rises, the squares in the attribution results become more complete, rather than just edges or noise.

C Experimental Results of TIF

C.1 Class Sensitivity

Figure 12 show the intuitive results of class sensitivity of TIF(α=0.7) on gradient-based attributions. It can be seen that different methods have different class sensitivities, but all have achieved completeness gains.

Figure 13 provide the TIF(α=0.7) of backdoor attack, which have the similar results.

C.2 Fine-grained Localization

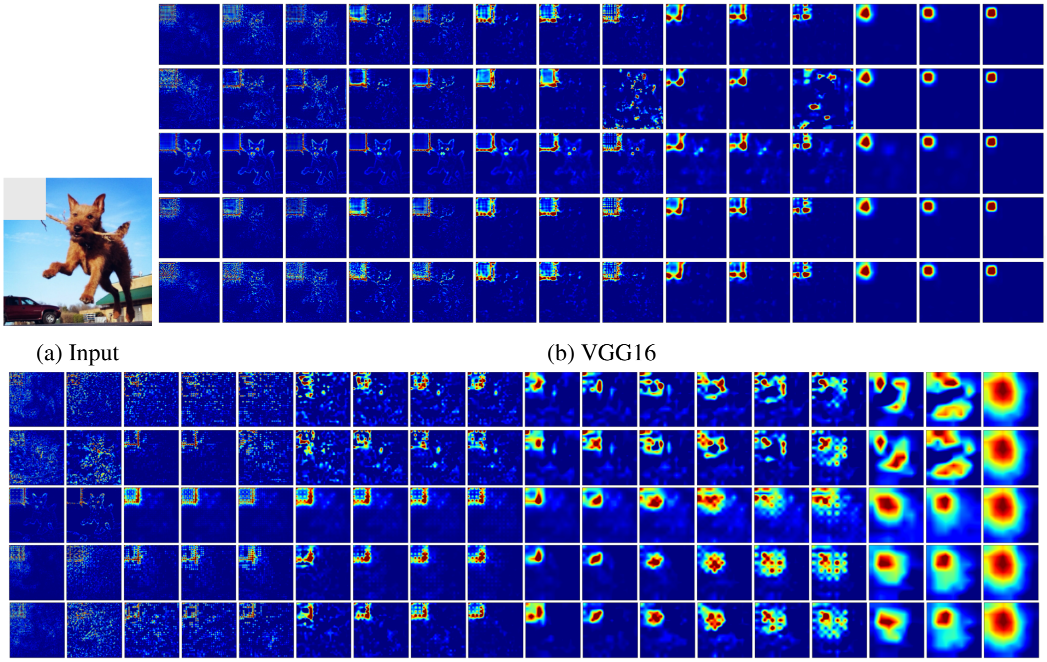
Figure 14 show the intuitive results of middle layer attributions for cross trigger. The methods improved by TIF are successfully highlight the top left trigger X.

C.3 Completeness

Figure 15 show the intuitive results of middle layer attributions for square trigger. The methods improved by TIF are successfully provide a complete square trigger.

C.4 Parameter Analysis

Figure 16 and 17 provide the intuitive influence of α, it can be seen that the larger the α, the better the fine-grained localization, but the worse the completeness.

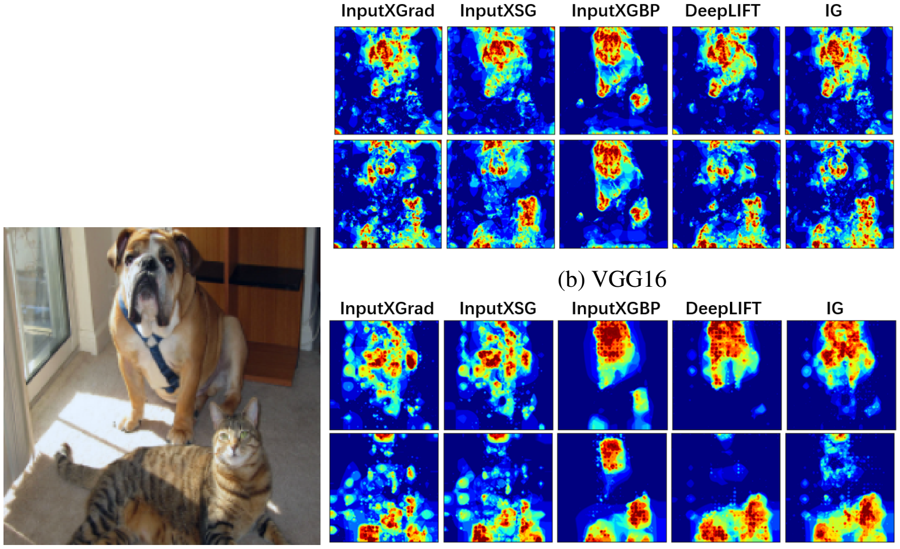


(a) Input

(b) VGG16

(c) ResNet50

Figure 11: Middle layer attributions for InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient with backdoor triggers of completeness.



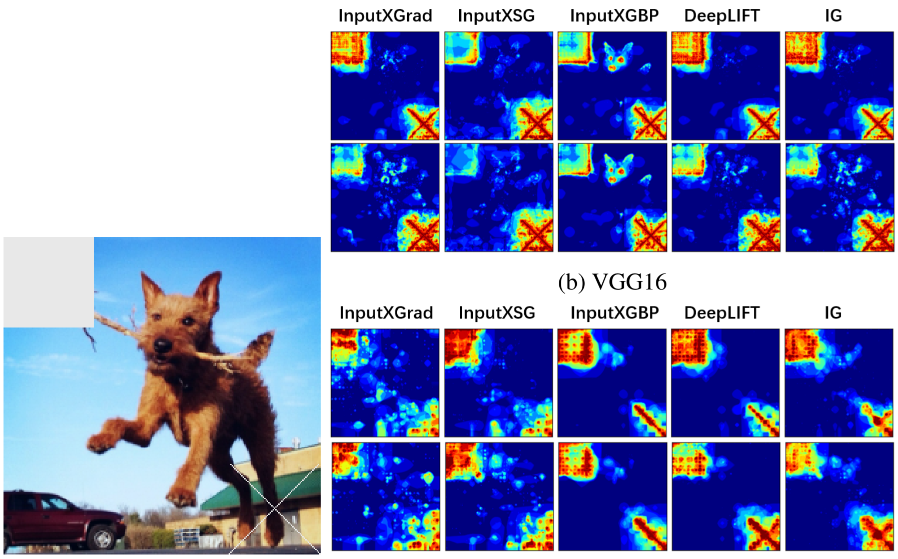
(b) VGG16

(a) Input

(c) ResNet50

Figure 12: Intuitive results for TIF(0.7) applyed to InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient.

Figure 18 demonstrate how α influence the performance on reliability metrics. All α brings enhancements, and different α shows a trade-off between LeRF and MoRF.

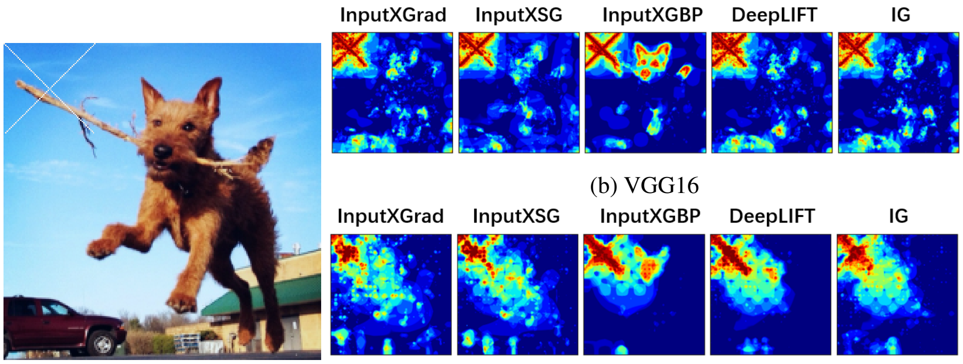


(b) VGG16

(a) Input

(c) ResNet50

Figure 13: Intuitive results for TIF(0.7) applyed to InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient with backdoor triggers of class sensitivity.

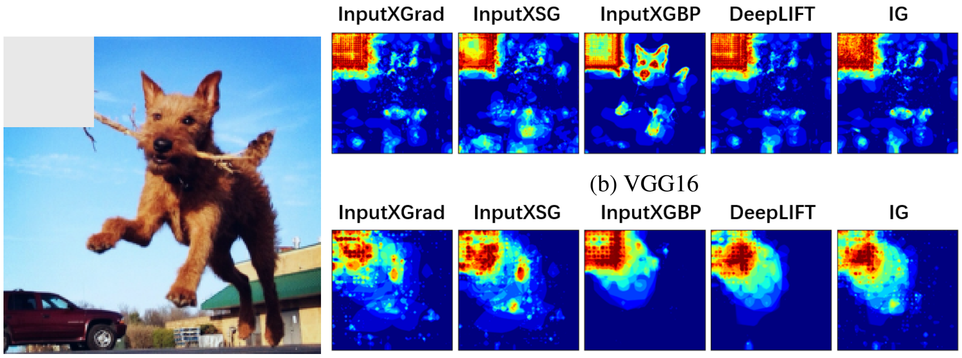


(b) VGG16

(a) Input

(c) ResNet50

Figure 14: Intuitive results for TIF(0.7) applyed to InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient with backdoor triggers of fine-grained localizations.

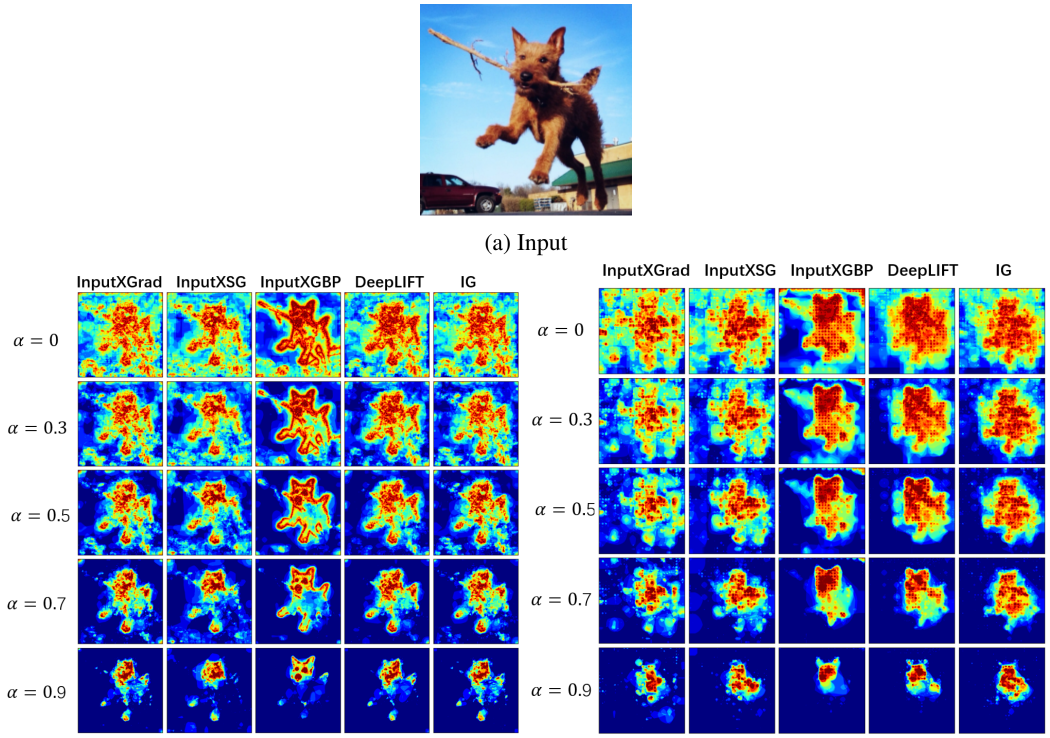


(b) VGG16

(a) Input

(c) ResNet50

Figure 15: Intuitive results for TIF(0.7) applyed to InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient with backdoor triggers of completeness.

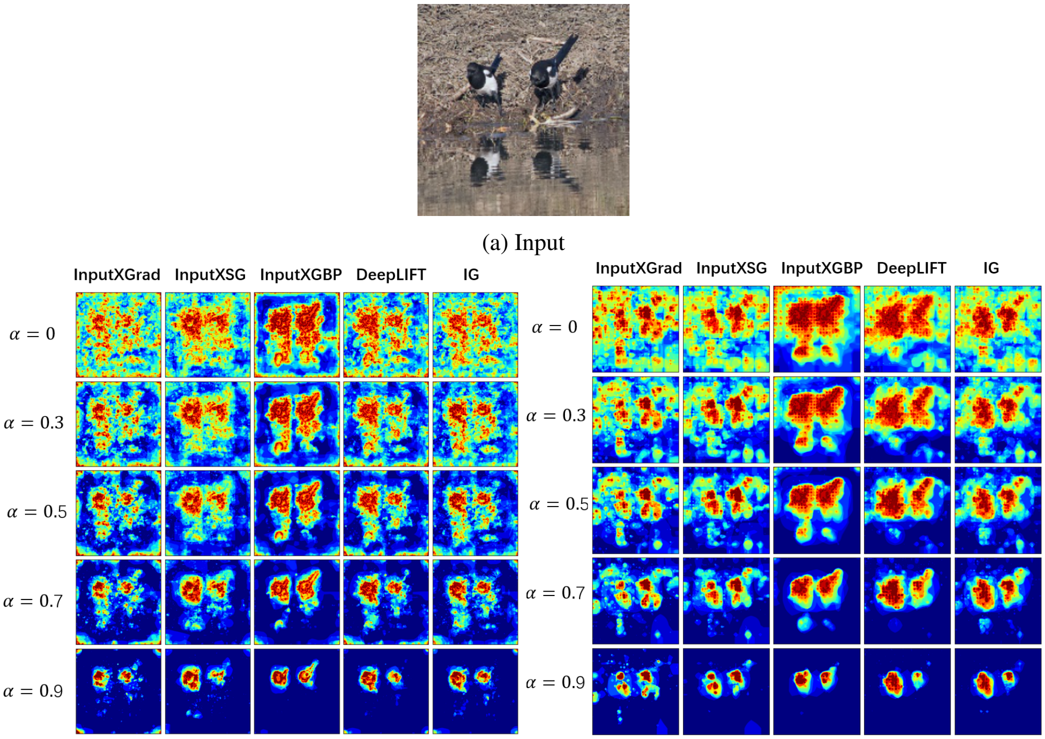


(a) Input

(b) VGG16

(c) ResNet50

Figure 16: Intuitive results for TIF applied to InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient with different α.



(a) Input

(b) VGG16

(c) ResNet50

Figure 17: Intuitive results for TIF applyed to InputXGrad, InputXSG, InputXGBP, DeepLIFT, IntegratedGradient with different α.

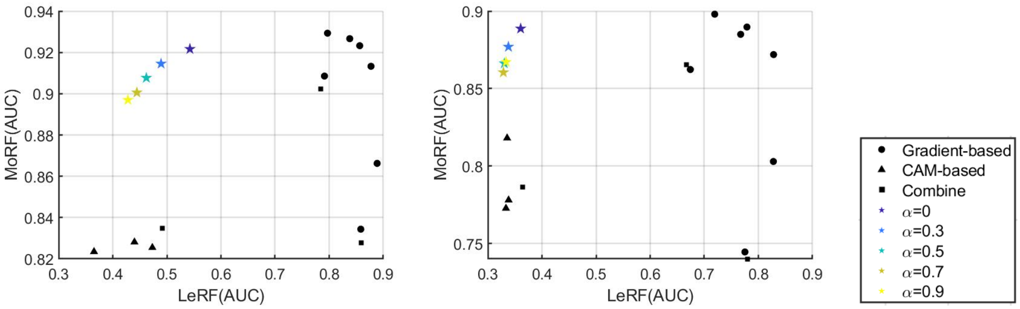


Figure 18: Parameter analysisi of α for pixel perturbation. The left one is result on VGG16 and the right one is ResNet50. α shows a trade-off between LeRF and MoRF.