Generating Deep Networks Explanations with Robust Attribution Alignment

Guohang Zeng GUOHANGZ@STUDENT.UNIMELB.EDU.AU
Yousef Kowsar KOWSAR.YOUSEF@UNIMELB.EDU.AU
Sarah Erfani SARAH.ERFANI@UNIMELB.EDU.AU
James Bailey BAILEYJ@UNIMELB.EDU.AU

School of Computing and Information Systems
The University of Melbourne, Australia

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Attribution methods play a key role in generating post-hoc explanations on pre-trained models, however it has been shown that existing methods yield unfaithful and noisy explanations. In this paper, we propose a new paradigm of attribution method: we treat the model's explanations as a part of network's outputs then generate attribution maps from the underlying deep network. The generated attribution maps are up-sampled from the last convolutional layer of the network to obtain localization information about the target to be explained. Inspired by recent studies that showed adversarially robust models' saliency map aligns well with human perception, we utilize attribution maps from the robust model to supervise the learned attributions. Our proposed method can produce visually plausible explanations along with the prediction in inference phase. Experiments on real datasets show that our proposed method yields more faithful explanations than post-hoc attribution methods with lighter computational costs.

Keywords: interpretability, adversarial robustness, deep neural networks, explanation generation, attribution method

1. Introduction

Deep neural networks (DNNs) have achieved great success in many areas, but they are also known as "black boxes". Recent studies on Interpretable Machine Learning (IML) (Ancona et al., 2017; Shrikumar et al., 2016; Zeiler and Fergus, 2014) aim at explaining deep networks and try to open the black-boxes of DNNs. Among various IML approaches, attribution methods (Selvaraju et al., 2017; Smilkov et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017; Simonyan et al., 2013) are well adopted approaches to explain DNNs. Given a deep network, attribution methods produce attribution maps which link the input data to their corresponding contribution to the model's prediction. This explanation is necessary to human users to help them understand how black-box models make decisions. For example, in healthcare domain, attribution methods are used to highlight saliency region of medical images to help physicians understand what features have been discovered by deep learning-based diagnosis systems (Singh et al., 2020).

Existing attribution methods make post-hoc explanations on pre-trained models. Given a pre-trained model, gradient-based attribution methods use the gradients of deep networks to explain its prediction. For example, deep saliency method (Simonyan et al., 2013) computes the gradients of the output with respect to the input to generate attribution map. Intuitively, gradients can represent the direction that has large impacts on the model's response, as it can theoretically represent the contributions of inputs for linear models. On the other hand, perturbation-based methods (Zeiler and Fergus, 2014) produce attribution maps by generating perturbations on input data that cause significant performance degradation. Since perturbations can cause degradation in model's performance, the magnitude of perturbations corresponds to the feature importance to the model. In addition to these methods, there are many variants of attribution methods (Lundberg and Lee, 2017) (Sundararajan et al., 2017) (Shrikumar et al., 2016) (Springenberg et al., 2014) (Fong and Vedaldi, 2018) (Yang et al., 2020) which provide explanations for DNNs.

However, most of attribution methods are heuristic, and generate attribution maps that are often noisy and incomprehensible to the humans. A number of studies have shown that attribution methods do not seem to provide faithful explanation for the model. (Hooker et al., 2018) empirically showed that most of attribution methods often generate unfaithful explanations that are not even better than random designation of feature importance. More surprisingly, existing attribution methods seem to be independent of the model (Adebayo et al., 2018).

1.1. Motivation of this work

Most of the previous interpretability methods are post-hoc methods that explain a pre-trained machine learning model. However, this approach has two significant drawbacks: (a) interpretability methods are independent of models, making the explanations and the models less relevant. (b) most of the post-hoc interpretability methods utilize the backpropagation rule to pass feature importance from the output to the input layer, thus suffered from backpropagation-related issues such as gradient saturation (Sundararajan et al., 2017).

In this paper, we aim to improve the faithfulness of interpretability methods, i.e., making attribution maps to be indicative of true feature importance. Since post-hoc explanations can be independent of the model (Adebayo et al., 2018), we consider a different paradigm of attribution method to overcome the above limitations—we treat attribution method as a part of neural network's output in order to increase the correlation between the attribution map and the underlying deep network. By coupling these two components as shown in Figure 1, our method can generate end-to-end attribution maps by a branch network called *explainer*. The generated attribution maps are up-sampled by the explainer from the last convolution layers, which are considered to be able to localize discriminative image regions (Zhou et al., 2016) for the prediction. Our proposed method can be regarded as a variant of class activation maps(CAMs) methods (Selvaraju et al., 2017) (Zhou et al., 2016) that uses the features on the last convolutional layers to visualize DNNs. Unlike previous CAMs methods using bilinear-interpolation to upsample on a weighted mean feature maps, we leverage the explainer network to learn a non-linear mapping from the last convolutional layer to generate the attribution map, while preserving all channel information in the last convolutional layers.

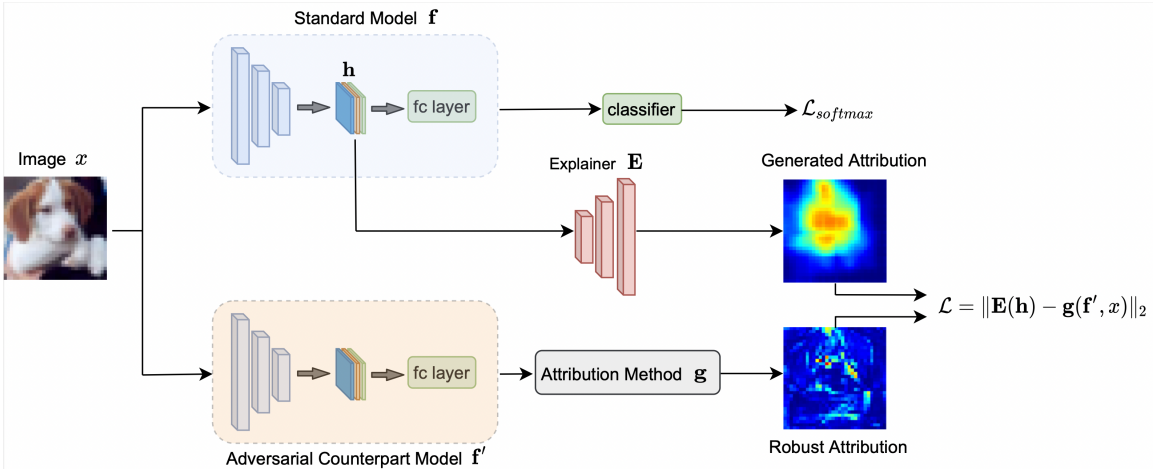

Figure 1: An overview of training phase of our proposed method. The generated attribution map is up-sampled from the last convolutional layer. Robust attribution from a pretrained adversarially counterpart model is used to align the generated attribution.

Another intuition of our work lies on a connection of adversarial robustness and interpretability, which has been observed by many studies (Moshkovitz et al., 2021; Etmann et al., 2019; Tsipras et al., 2018; Zhang and Zhu, 2019). Specifically, attribution maps from adversarially robust model have shown to align well with human perception (Tsipras et al., 2018), and are more biased towards shape-based features and alleviates texture-based features (Zhang and Zhu, 2019). To this end, we utilize robust model's attribution (which we call *robust attribution* in this paper) to align the generated attribution map to learn visually comprehensible explanations. In the inference phase, our method generates attribution map by forwarding the input to the explainer network, which has lighter computational cost compared with other post-hoc attribution methods.

In summary, our study makes the following contributions:

- We propose a new paradigm of attribution method, to the best of our knowledge, this is the first approach coupling attribution map with the underlying model. We use an explainer network to up-sample features from the last convolutional layer to learn localization information about the object to be explained. Our method can generate end-to-end attributional explanation with lighter computational costs.
- We conduct quantitative evaluations on the attribution maps from adversarially robust models and show that robust models can yield more faithful attribution maps, which can be used to align the generated attribution map.
- By extensive evaluation on CIFAR10 and TinyImageNet datasets, we show that our method not only produces clearer and human-comprehensible explanations, it also produces faithful attribution maps that contribute to the model's prediction.

2. Related Work

In this section we give review of two groups of related studies: (1) attribution methods; (2) benefits of adversarially robust models.

2.1. Attribution Methods

Gradient-based attribution methods have been widely adopted to generate attribution maps for explaining DNNs. (Simonyan et al., 2013) computes the gradients of the output with respect to the input to generate saliency maps. To enhance the quality of saliency map, several variant saliency methods have been proposed. GradientXinput (Shrikumar et al., 2016) method sharpens attribution maps via element-wise multiplication between the input and its gradients. Guided backpropagation (Springenberg et al., 2014) only considers positive gradients to represent the contribution of input instances, and ignoring negative gradients when back-propagating. Integrated Gradients (Sundararajan et al., 2017) assigns attribution map by computing the integral of gradients of the model's outputs respect to inputs along the path from a chosen baseline to inputs. DeepLIFT (Shrikumar et al., 2017) designs hand-crafted propagation rules to replace chain rule when propagating the output of network to its input.

On the other hand, perturbation-based methods generate attribution maps by altering input features, then measure the change of the model's output to represents the important weights for the modified input. (Zeiler and Fergus, 2014) proposed a method to occlude different regions of the image then computes attribution maps by the change of output. Perturbation-based methods are generally much slower than gradient-based methods since they need to inference the model multiple times.

Despite the above methods, SmoothGrad (Smilkov et al., 2017) is an ensembling approach that can be used to incorporate with base attribution methods to improve the quality of the base methods. SHAP (Lundberg and Lee, 2017) methods use the shapley values (Shapley, 1953) to explain model's predictions. LIME (Ribeiro et al., 2016) generates attribution maps by locally approximates a nonlinear model with a linear function on the input instance. Class activation map methods (Selvaraju et al., 2017) (Zhou et al., 2016) can be regarded as special cases of gradient-based methods that only utilized the last convolutional layers to generate attribution maps. (Simpson et al., 2019) penalizes attribution maps when they are not consistent with the lesion segmentation to reduce overfitting in medical imaging.

2.2. Benefits of Adversarially Robust Model

Deep networks have shown to be vulnerable to adversarial attacks (Szegedy et al., 2013) (Goodfellow et al., 2014), which are crafted by adding imperceptible perturbations on input data to cause false predictions. To enhance model's robustness to adversarial attacks, adversarial training (Goodfellow et al., 2014) (Madry et al., 2017) is the current state-of-the-art defense technique that augments the training data with adversarial examples when training the model. Recent studies (Engstrom et al., 2019; Tsipras et al., 2018; Salman et al., 2020; Singh et al., 2019; Zhang and Zhu, 2019) showed that adversarially trained models

are not only resistant to adversarial attacks, but also have some unexpected benefits and have connections to interpretability (Etmann et al., 2019).

(Engstrom et al., 2019) demonstrated that adversarially robust training can be viewed as including a human prior over the representation that models are able to learn. (Tsipras et al., 2018) showed that the saliency maps of adversarially robust models align well with human perception. (Etmann et al., 2019) proved that the alignment between input image and saliency map is positively correlated with adversarial robustness. By extensive experiments, (Zhang and Zhu, 2019) demonstrated that the representation learned by robust models are more biased towards shape-based features and alleviates texture-based features. To this end, (Singh et al., 2019) showed that attribution map from adversarially robust model are also robust to perturbations, then utilized these characteristics of robust attribution on weakly supervised object localization.

3. Generating Explanations with Robust Attribution Alignment

In this paper, we propose a new paradigm of attribution method which incorporates attribution with neural network architecture. We utilize robust model's attribution to train the proposed method and learn visually comprehensible explanations, which can be seen as another benefits of adversarially robust model.

3.1. Problem Definition

Consider a machine learning model $\mathbf{f} : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$, where $x \in \mathbb{R}^d$ denotes the input data, and $\mathbf{f}(x) \in \mathbb{R}^{d'}$ denotes its prediction. A post-hoc attribution method is a function $\mathbf{g} : \mathcal{F} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ that provides attribution map (i.e., feature importance estimation) $\mathbf{g}(\mathbf{f}, x)$ for the given model and its prediction on test instance x. In this paper, we consider a different paradigm, where the attribution map $\mathbf{g}(\mathbf{f}, x)$ is an output of the model $\mathbf{f}$. Specifically, we aim to train a machine learning model $\mathbf{f} : \mathbb{R}^d \mapsto \mathbb{R}^{d+d'}$ where the output of the model contains its original output $y = \mathbf{f}(x) \in \mathbb{R}^{d'}$ along with feature importance estimation $\mathbf{g}(\mathbf{f}, x) \in \mathbb{R}^d$.

3.2. Our Approach

As shown in Figure 1, our proposed method treats the task of generating attribution map as a part of the network's output. Given a deep network $\mathbf{f}$, we train a branch network called *explainer*, which upsamples feature maps of the last convolutional layers to generate attribution maps. Inspired by recent studies that showed adversarially robust models can generate more visually comprehensible saliency map (Tsipras et al., 2018), we utilized the robustly trained model's attributions to supervise the generated attribution map in order to learn visually comprehensible attribution maps.

There are three steps to train our proposed model. First, we train an adversarially robust model with identical network architecture as $\mathbf{f}$, refered as the adversarially robust counterpart model $\mathbf{f}'$. Next, we use the counterpart model $\mathbf{f}'$ to generate robust attributions. In the final step, we train the explainer network supervised by aligning the outputs of the explainer to the robust attributions.

3.2.1. Training Adversarially Robust Counterpart Model

Given a model $\mathbf{f}$, we first train an adversarially robust counterpart model $\mathbf{f}'$ with the identical network architecture. The counterpart model $\mathbf{f}'$ only serves to generate robust attribution maps rather than defend adversarial attacks. In this work, we use adversarial training to obtain model $\mathbf{f}'$.

Adversarial training aims to improve the robustness of a network by training with adversarial samples, which can be formulated as the following robust optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right] \tag{1}$$

where δ denotes adversarial perturbation, L is the loss function, $\mathcal{D}$ is the distribution of training data and $x' = x + \delta$ denotes adversarial example. In this paper we adopt projected gradient descent (PGD) (Madry et al., 2017) to generate adversarial examples, since we can control the degree of robustness of the model by controlling the predefined range of PGD adversary. PGD method iteratively computes an adversarial example as:

$$x'_{t+1} = \Pi_{clip}\left(x'_t + \alpha \operatorname{sgn}\left(\nabla_x L(\theta, x, y)\right)\right) \tag{2}$$

where α is attack learning rate and Π_{clip} is a clip function maintaining x' within a predefined range $\|x' - x\|_p \leq \varepsilon$. Specifically, we use ℓ_2-bounded attacks to generate adversarial example x', thus the clip function $\Pi_{\text{clip},\varepsilon}$ is computed as:

$$\Pi_{\text{clip},\varepsilon}(x) = \begin{cases} x + \frac{x'-x}{\|x'-x\|_2}\varepsilon & \text{if } \|x'-x\|_2 > \varepsilon \\ x' & \text{otherwise} \end{cases} \tag{3}$$

Adversarial examples are crafted by Equation 2 and Equation 3. We train the robust model by optimizing the min-max problem as shown in Equation 1. In this work, we adopt PGD adversarial training because the saliency map of PGD-trained models are shown to align well with human perception (Tsipras et al., 2018). We use the robust model $\mathbf{f}'$ to generate the attributions map. As discussed in (Singh et al., 2019), attribution maps from robust models are also robust towards perturbations, hence we use robust attributions generated from these models.

3.2.2. Robust Attribution

In this stage, we obtain the robust attribution from the pretrained adversarial counterpart model $\mathbf{f}'$. As shown by (Tsipras et al., 2018), saliency maps of robust models align well with human perception. To this end, we choose saliency map to generate robust attributions.

Let $S(x) = \frac{\partial A^c}{\partial x}$ denote the saliency map of x, and A^c is the output activation of interest. To improve the quality of saliency map, we use SmoothGrad as an ensembling method to denoise saliency map of robust model. SmoothGrad uses a set of Gaussian noise to perturb a given input x then averages the attributions maps from the set of perturbed inputs. Formally, the robust attribution is computed as:

$$\mathbf{g}(\mathbf{f}', \mathbf{x}) = \frac{1}{n}\sum^{n} S\left(x_i + \mathcal{N}\left(0, \sigma^2\right)\right) \tag{4}$$

where n is the number of perturbed data and $\mathcal{N}\left(0, \sigma^2\right)$ denotes the Gaussian noise used to perturb the input data.

3.2.3. Training Explainer with Robust Attributions

After we obtain robust attributions from robust counterpart model $\mathbf{f}'$, we utilized these robust attributions to train our explainer network. As discussed by (Zhou et al., 2016) (Selvaraju et al., 2017), the high-level features of deep convolution layers is able to localize the object region in the image. To this end, we utilized the feature maps of the last convolutional layers to learn the generated attribution maps.

Let $\mathbf{h} \in \mathbb{R}^{C \times H \times W}$ denote the feature maps from the last convolutional layers, where C, H, W are their channel numbers, height and width, respectively. Explainer network is a function $\mathbf{E} : \mathbb{R}^{C \times H \times W} \mapsto \mathbb{R}^d$ maps $\mathbf{h}$ to the input data space which has the same size as attribution maps. Different from CAMs methods that squeeze C channels into 1 channel by taking mean of weighted activation maps of $\mathbf{h}$, we utilize all channel information in $\mathbf{h}$ to generate attribution map. Specifically, we adopt sequential deconvolutional layers with batch normalization to be the explainer network.

In addition to obtaining localization information from high-level features $\mathbf{h}$, we let the generated attribution maps learn visually comprehensible features from the robust attributions. We align the generated attribution maps with robust attributions by minimizing their ℓ_2 distance. Formally, the explainer network is trained with the original task of model $\mathbf{f}$:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[L(\theta, x, y) + \lambda \|\mathbf{E}(\mathbf{h}) - \mathbf{g}(\mathbf{f}', x)\|_2 \right] \tag{5}$$

where λ is the loss weight for balancing, and $L(\theta, x, y)$ is the cross-entropy loss. After the model is trained, we can obtain attribution map by forward input data x from $\mathbf{h}$ to the explainer network. Our approach can generate end-to-end attribution maps with lighter computational costs compared with post-hoc attribution methods.

4. Experiments

In this section, we evaluate the performance of our proposed method in terms of two evaluation protocols in comparison with post-hoc attribution methods. We describe the datasets we used, implementation details, evaluation protocols and compared methods, respectively. Then we show the experiment results over different evaluation protocols.

4.1. Datasets and Implementation Details

We conduct experiments on two datasets: CIFAR10 and TinyImageNet.

- **CIFAR10**: It is a standard vision dataset containing 60,000 natural images with low resolution.
- **TinyImageNet**: It is a tiny version of ImageNet. It has 200 classes of objects. Each class has 500 training images, 50 validation images, and 50 test images. Since the label of testsets is not open to the public, we choose validation sets as testsets in our experiment settings.

For both datasets, we use the ResNet-18 (He et al., 2016) model architecture. We choose ResNet-18 because it gives good performance in a reasonable amount of training time. In the data augmentation stage, we use zero paddings with 4 width on images, then perform random horizontal flip and random crop.

The explainer network is consists of 4 consecutive deconvolutional layers followed by batch normalization layers. Each filter has a stride of 3, the number of filters in the 4 deconvolutional layers are 64, 128, 256, 512, respectively. ReLU is the activation function used for all of the layers. The loss weight λ is set to 0.35.

In the training phase, we use SGD optimizer with momentum at 0.9, learning rate is initially set to 0.1 with step-wise learning rate decay schedule. The batch size is set to 256 for CIFAR10 and 128 for TinyImageNet. As for the PGD adversarial training, we use ℓ_2-bounded attacks to generate PGD adversary x'. We performed a grid search over the hyper-parameters ϵ and iterations size then found that $\epsilon = 1.5$ and iterations size $= 20$ are the hyper-parameters to yield the best results.

4.2. Evaluation protocols

In this paper, we use two protocols to evaluate the performance of attribution map: Remove and Retrain and Sanity Check.

4.2.1. Remove and Retrain (ROAR)

It is crucial to quantitatively measure whether the attribution maps are in fact contributing important features for the model's prediction. ROAR (Hooker et al., 2018) measures the performance of attribution map by removing features from the input then looks at how the classifier degrades. As discussed by (Hooker et al., 2018), a previous similar strategy (Samek et al., 2016) induced distribution shift to training and evaluation data. To this end, ROAR trains the model when a subset of the features are removed. Specifically, ROAR first replaces fraction of the pixels considered to be important features with a constant mean value. Then, the model is retrained and test on the modified dataset. The intuition behinds ROAR is that a faithful attribution map should lead to great degradation of model's performance. ROAR requires an extremely time consuming evaluation metric due to the need to retrain the network multiple times.

4.2.2. Sanity Check

(Adebayo et al., 2018) proposed the sanity check method to check if attribution maps look different when the deep network being explained is extremely perturbed. The intuition behind this measure is that a faithful attribution method should yield different explanation on the randomized model. Surprisingly, (Adebayo et al., 2018) has shown that most of attribution maps do not pass the sanity check—they are visually indistinguishable before and after we randomize the weights of the network.

4.3. Compared Methods

We compare our proposed method with the following methods:

i. **Random**: It assigns feature importance weights randomly from a uniform distribution from 0 to 1. The random baseline tells us how much better is an attribution method compared to a random designation of feature importance.

ii. **Saliency (Simonyan et al., 2013)**: It computes the gradients of the output with respect to the input to generate attribution maps.

iii. **Intergrated Gradients (IG) (Sundararajan et al., 2017):** It computes the average gradients of multiple inputs by introducing integration paths.

iv. **DeepLIFT (Shrikumar et al., 2017)**: It backpropagates the output with respect to the input via hand-crafted propagation rules.

v. **Gradients Shap(GradShap) (Lundberg and Lee, 2017):** It approximates shapley values by computing the expectations of gradients by randomly sampling from chosen baselines.

vi. **Occlusion (Zeiler and Fergus, 2014) :** It occludes different regions of the image then generate attribution maps by the change of output.

vii. **SmoothGrad-Saliency (SG-S) (Smilkov et al., 2017):** SG-S use a set of Gaussian noise to perturb a given input x then averages the base attributions maps from the set of perturbed inputs. Here we adopt Saliency maps as base attribution maps.

4.4. Experiment Results

4.4.1. Implementation details

For DeepLIFT and Intergrated Gradients, baselines are set to zero. As for the hyper-parameters of Occlusion method, the sliding window shapes is set to (3,4,4) and the strides is set to 2. The sampling size is set to 50 for SmoothGrad method. For GradientShap method, baselines are generated from a set of Gaussian distributions with $\sigma = 0.001$. We also compare the above method to the robust attribution, which is denotes as *Robust Attribution* in the following experiments. We implemented the above attribution methods by the captum library (Kokhlikyan et al., 2020).

4.4.2. Visualization

Figure 2 shows the visualization of our methods and the attribution methods on CIFAR10 dataset with ResNet-18 model. As shown in the figure, our method is more visually comprehensible and is able to cover discriminative regions for the object in images.

4.4.3. Evaluation of ROAR

Table 1 and Figure 3 shows a comparative evaluation of the proposed method with other compared methods in terms of ROAR measures on CIFAR10 and TinyImageNet. We generate modified datasets at different remove rate $\eta = [20\%, 40\%, 60\%, 80\%]$. Afterwards the model is re-trained on the modified dataset and evaluated on the new test dataset. Lower accuracy indicates the attribution method is more faithful to the model. As shown in the

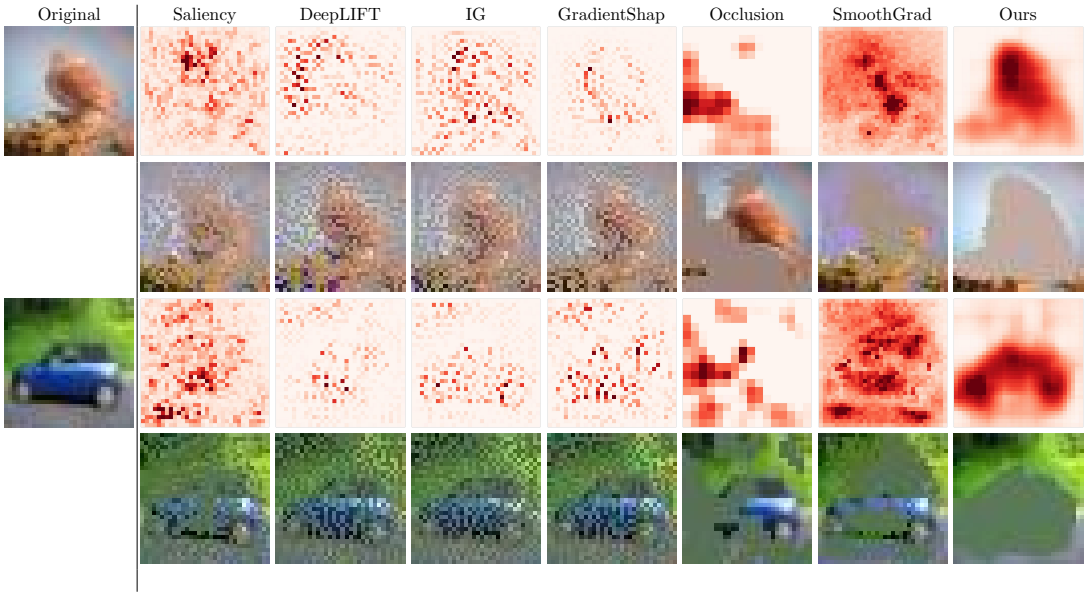

Figure 2: Visualization of our methods against post-hoc attribution methods on ResNet-18 (CIFAR10). The even rows show images when top 50% fraction of pixels estimated to be most important by each attribution methods is replaced with the mean.

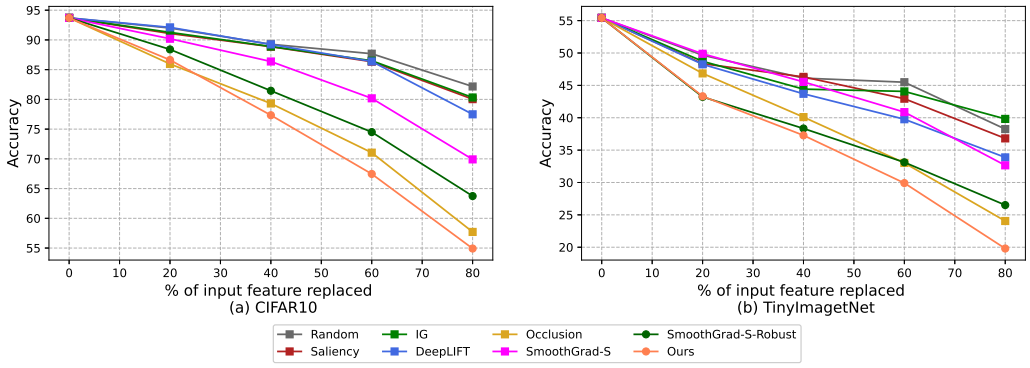

Figure 3: ROAR evluation on CIFAR10 and TinyImageNet datasets with the ResNet-model. Top 20, 40, 60, 80% of important pixels of each image are replaced with a constant mean value. The degree of performance degradation of the model after being retrained on the new dataset indicates the faithfulness of each attribution method. Our method outperforms than other attribution methods. (lower is better)

figure, our method has better ROAR performance especially when η is large. The Occlusion method has the closest performance to our method, but it has much higher computational cost (see discussion in section 4.4.7). We also found that the robust attribution is a more faithful explanation than its standard counterpart, since robust attribution is less noisy (Tsipras et al., 2018) and focus more on shape information (Zhang and Zhu, 2019).

Table 1: Overall Remove and Retrain(ROAR) evaluation for all compared attribution methods and our proposed method.

Datasets	Methods	Remove and Retrain Rate η			
		20%	40%	60%	80%
CIFAR10	Random	92%	89.27%	87.68%	82.17%
	Saliency	91.08%	88.86%	86.34%	80.04%
	Integrated Gradients	91.27%	88.89%	86.5%	80.32%
	DeepLIFT	92.11%	89.24%	86.38%	77.49%
	GradShap	89.7%	87.59%	85.7%	78.89%
	Occlusion	**85.96%**	79.31%	71.04%	57.71%
	SmoothGrad	90.2%	86.37%	80.18%	69.91%
	Robust Attribution	88.4%	81.45%	74.50%	63.74%
	ours	86.63%	**77.34%**	**67.46%**	**54.92%**
TinyImageNet	Random	49.66%	46.12%	45.49%	38.25%
	Saliency	48.35%	46.31%	42.94%	36.81%
	Integrated Gradients	48.72%	44.42%	44.07%	39.82%
	DeepLIFT	48.25%	43.7%	39.77%	33.88%
	GradShap	48.09%	44.11%	40.74%	35.36%
	Occlusion	46.85%	40.1%	33.02%	24.05%
	SmoothGrad	49.88%	45.54%	40.87%	32.65%
	Robust Attribution	**43.22%**	38.34%	33.12%	26.51%
	ours	43.32%	**37.28%**	**29.91%**	**19.8%**

4.4.4. Explaining model $\mathbf{f}$ by attribution maps from counterpart model $\mathbf{f}'$

In this section, we investigate whether the robust attribution can be used to explain its standard counterpart model $\mathbf{f}$. Previous studies have shown that attribution maps of robustly trained models align well with human perception, we show that they also yield faithful explanation in terms of ROAR evaluation. We use the robust attribution from the adversarially trained counterpart model $\mathbf{f}'$ to indicate the attribution map for $\mathbf{f}$. Figure 4 shows a comparative ROAR evaluation of robust attribution with standard attribution. As shown in the figure, attribution maps from robustly trained model $\mathbf{f}'$ are consistently better than the original attribution maps. The empirical experiment shows that adversarially robust model can be used to yield more faithful explanation for its counterpart standard model, which can be seen as another previously undiscovered benefit of adversarially robust model. Both ℓ_2-trained models and ℓ_{inf}-trained models have this effect. Due to the space constraints, we include more experiments results in the supplementary material.

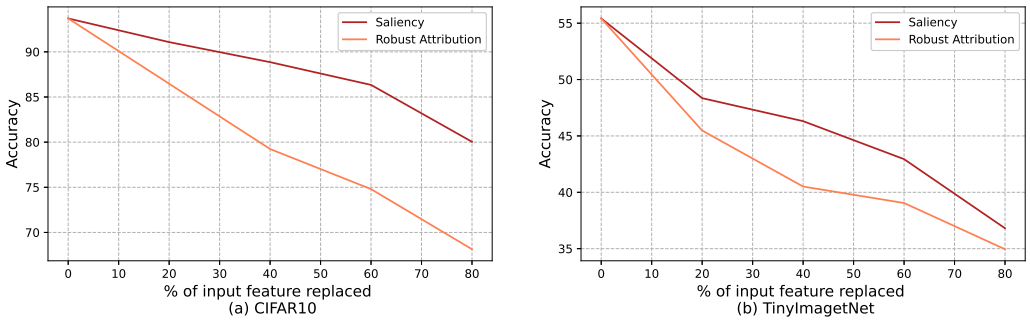

Figure 4: ROAR comparison between saliency of standard model and saliency of robust model. Adversarial counterpart model can be used to generate more faithful attributions.. (lower is better)

4.4.5. Assessing Faithlessness of attribution

In this section, we conduct sanity check to evaluate the faithfulness of attribution methods– if an attribution map is faithful, then it should be different when the weights of the network are randomized. In Figure 5, we demonstrate visualizations of various attribution maps on the original ResNet-18 model and a perturbed ResNet-18 model where all the convolutional layers in the last Residual Block are replaced to a set of zero-mean Gaussian noise $\mathcal{N}\left(0, \sigma^2\right)$. A faithful attribution map should have a different visualization result, since the weight of the network has been completely perturbed. The prediction of the randomized model degrades to random guess.

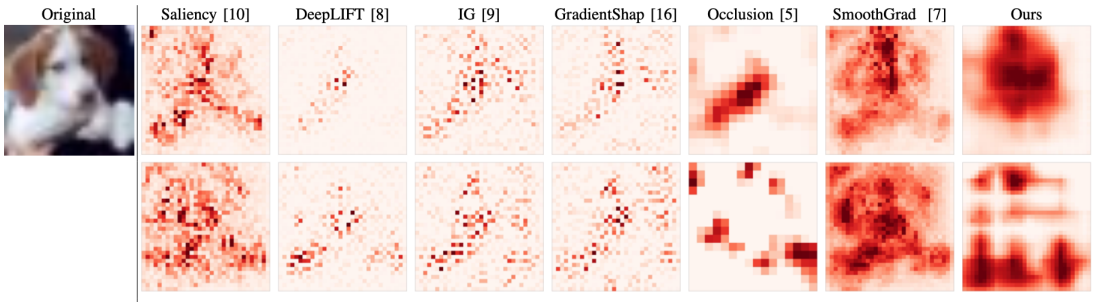

Figure 5: Sanity check: examples of various attribution maps for the original model(first row) and the randomized model(second row)

As we see in the figure: saliency, DeepLIFT and Integrated Gradients didn't pass the sanity check—they look similar when the network being explained is extremely perturbed. The attribution maps of our method for the two model look distinguishable. This is due to the fact that our method is generated directly on the feature map of the model, thus it is sensitive to perturbations acting on the model.

4.4.6. Ablation study

In this section, we conduct ablation study to analyze whether the generated explanations can benefit from robust attribution alignment. We replaced the robust model with the standard model then used SmoothGrad Saliency from standard model to train the explainer network, denoted as *standard alignment*. As shown in Table 2 and Figure 6, attributions that learned with standard attribution alignment can not generate visually plausible attribution map, leading to ROAR performance degradation (its ROAR performance is near to random feature importance assignation). This confirm that robust attribution does help the explainer network to learn better attribution maps.

Table 2: Ablation study: impact of robust attribution alignment

Datasets	Methods	Remove and Retrain Rate η			
		20%	40%	60%	80%
CIFAR10	Explainer + Standard Alignment	91.18%	88.24%	86.68%	81.47%
	Explainer + Robust Alignment	86.63%	77.34%	67.46%	54.92%
TinyImageNet	Explainer + Standard Alignment	48.64%	45.92%	43.27%	37.88%
	Explainer + Robust Alignment	43.32%	37.28%	29.91%	19.8%

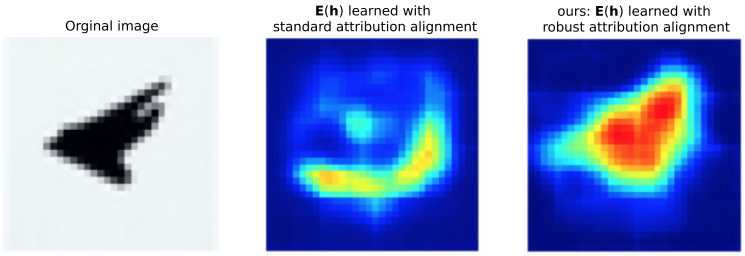

Figure 6: Ablation Study: from left to right: original image, heatmap of generated attribution with standard attribution alignment, ours.

4.4.7. Computational efficiency

We evaluate the computational efficiency of the proposed method and all compared attribution methods based on ResNet-18 network architecture. The empirical experiments is conducted on Pytorch CPU mode with Intel i7 CPU. Table 3 shows the computational costs of compared attribution methods and our proposed model. Our method has the best computational efficiency in inference phase over all the compared methods, since it generates explanations by forwarding input through the explainer network without backpropagating. It is noteworthy that Occlusion method

Table 3: Computational cost

Attribution Methods	Latency (ms)
Saliency	19
Integrated Gradient	1385
DeepLIFT	35
GradShap	84
Occlusion	1875
SmoothGrad	972
ours	**9**

has a close ROAR performance to our method, but its computational efficiency is almost 200 times lower than ours.

Our method requires additional workload to train the explainer network. As discussed in Section 3.2.3, we train the explainer network along with the given deep neural network to be explained, which induces additional workload in the training phrase. In other word, our method takes the workload in the training phrase to trade the faithfulness of explanations and computational efficiency in the inference phrase.

5. Conclusion

In this paper, we proposed a new paradigm of attribution method that enables the process of generating attribution maps as a part of deep network's output, to increase the correlation between the explanation and the underlying model. The generated attribution maps leverage the localization information from the last convolutional layer and the shape-based information from the adversarially trained counterpart model. This paper also points out an interesting application of adversarially robust model, as it can produce more faithful explanations to its standard counterpart model. As demonstrated on CIFAR10 and TinyImageNet datasets, our approach end-to-end produces more faithful explanation than widely used post-hoc attributions with lighter computational costs.

Acknowledgments

This research was supported by the Melbourne Research Scholarship. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne.

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.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, July 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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*, pages 770–778, 2016.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *arXiv preprint arXiv:1806.10758*, 2018.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Michal Moshkovitz, Yao-Yuan Yang, and Kamalika Chaudhuri. Connecting interpretability and robustness in decision trees through separation. *arXiv preprint arXiv:2102.07048*, 2021.

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 international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *arXiv preprint arXiv:2007.08489*, 2020.

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.

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*, pages 618–626, 2017.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317, 1953.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

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

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

Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Gradmask: Reduce overfitting by regularizing saliency. *arXiv preprint arXiv:1904.07478*, 2019.

Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

Mayank Singh, Nupur Kumari, Puneet Mangla, Abhishek Sinha, Vineeth N Balasubramanian, and Balaji Krishnamurthy. Attributional robustness training using input-gradient spatial alignment. *arXiv preprint arXiv:1911.13073*, 2019.

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

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

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

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Learning propagation rules for attribution map generation. In *European Conference on Computer Vision*. Springer, 2020.

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

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International Conference on Machine Learning*, pages 7502–7511. PMLR, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.