Enhanced Regularizers for Attributional Robustness

Anindya Sarkar, Anirban Sarkar, Vineeth N Balasubramanian
Indian Institute of Technology, Hyderabad
anindya.sarkar@cse.iith.ac.in, cs16resch11006@iith.ac.in, vineethnb@ith.ac.in

Abstract

Deep neural networks are the default choice of learning models for computer vision tasks. Extensive work has been carried out in recent years on explaining deep models for vision tasks such as classification. However, recent work has shown that it is possible for these models to produce substantially different attribution maps even when two very similar images are given to the network, raising serious questions about trustworthiness. To address this issue, we propose a robust attribution training strategy to improve attributional robustness of deep neural networks. Our method carefully analyzes the requirements for attributional robustness and introduces two new regularizers that preserve a model's attribution map during attacks. Our method surpasses state-of-the-art attributional robustness methods by a margin of approximately 3% to 9% in terms of attribution robustness measures on several datasets including MNIST, FMNIST, Flower and GTSRB.

Introduction

In order to deploy machine learning models in safety-critical applications like healthcare and autonomous driving, it is desirable that such models should reliably explain their predictions. As neural network architectures become increasingly complex, explanations of the model's prediction or behavior are even more important for developing trust and transparency to end users. For example, if a model predicts a given pathology image to be benign, then a doctor might be interested to investigate further to know what features or pixels in the image led the model to this classification.

Though there exist numerous efforts in the literature on constructing adversarially robust models (Qin et al. 2019; Wang et al. 2020; Zhang et al. 2019; Madry et al. 2018; Chan et al. 2020; Xie et al. 2019a), surprisingly very little work has been done in addressing issues in robustness of the explanations generated by a model. One aspect of genuineness of a model can be in producing very similar interpretations for two very similar human-indistinguishable images where model predictions are the same. (Ghorbani, Abid, and Zou 2019) demonstrated the possibilities to craft changes in an image which are imperceptible to a human, but can induce huge change in attribution maps without affecting the model's prediction. Hence building robust models, against such attacks proposed in (Ghorbani, Abid, and Zou 2019), is very important to increase faithfulness of such models to end users. Fig 1 visually explains the vulnerability of adversarial robust models against attribution-based attacks and how attributional training (à la adversarial training) addresses this to a certain extent. Such findings imply the need to explore effective strategies to improve attributional robustness.

The limited efforts until now for attributional training rely on minimizing change in attribution due to human imperceptible change in input (Chen et al. 2019) or maximizing similarity between input and attribution map (Kumari et al. 2019). We instead propose a new methodology for attributional robustness that is based on empirical observations. Our studies revealed that attributional robustness gets negatively affected when: (i) an input pixel has a high attribution for a negative class (non-ground truth class) during training; (ii) an attribution map corresponding to the positive or true class is uniformly distributed across the given image, instead of being localized on a few pixels; or and (iii) change of attribution, due to an human imperceptible change in input image (without changing predicted class label), is higher for a pixel with low attribution than for a pixel with high attribution (since this leads to significant changes in attribution). Based on these observations, we propose a new training procedure for attributional robustness that addresses each of these concerns and outperforms existing attributional training methods. Our methodology is inspired by the rigidity (and non-amorphous nature) of objects in images and instigates the fact that number of true class pixels are often small compared to total number of pixels in an image, resulting in a non-uniform (or skewed) pixel distribution of the true class attribution across spatial locations in the image. Complementarily, for the most confusing negative class, ensuring the attribution or saliency map is *not localized* helps indirectly improve the localization of the attribution of the true class.

The key contributions of our work can be summarized as follows: (i) We propose a *class attribution-based contrastive regularizer* for attributional training which forces the true class attribution to assume a skewed shape distribution, replicating the fact that few input pixels attribute highly compared to other pixels for a true class prediction. We also drive pixels of the attribution map corresponding to the negative class to behave uniformly (equivalent to not lo-

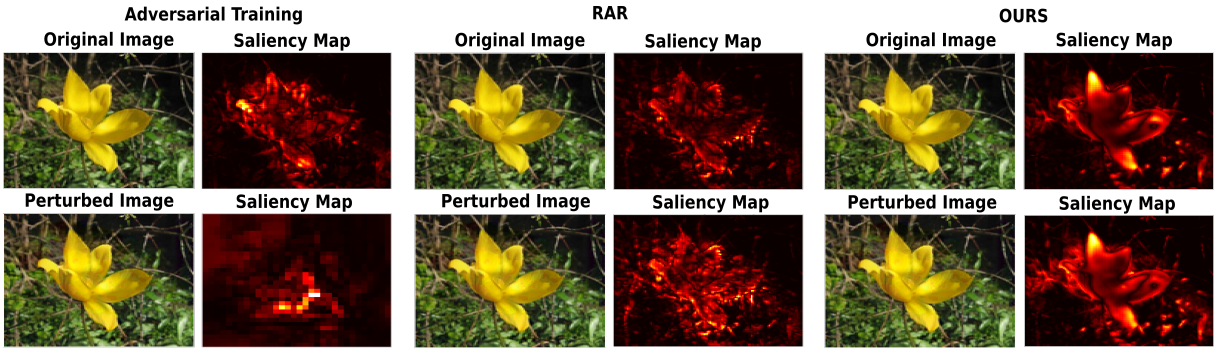

Figure 1: Original image and human-imperceptible attributional attacks (same predicted class) and their corresponding saliency maps under: *(Left)* adversarial training (Madry et al. 2018); *(Middle)* RAR (Chen et al. 2019); and *(Right)* our proposed robust attribution training for a test sample from Flower dataset. Note that the attacks/perturbed images in each column are different, since the attack is carried out based on the model trained on that particular method.

calizing) through this regularizer; (ii) We also introduce an *attribution change-based regularizer* to weight change in attribution of a pixel due to an indistinguishable change in the image (both the aforementioned contributions have not been attempted before to the best of our knowledge); (iii) We provide detailed experimental results of our method on different benchmark datasets, including MNIST, Fashion-MNIST, Flower and GTSRB, and obtain state-of-the-art results for attributional robustness across these datasets.

An illustrative preview of the effectiveness of our method is shown in Fig 1. The last column shows the utility of our method, where the attribution/saliency map (we use these terms interchangeably in this work) seems stronger, and minimally affected by the perturbation. The middle column shows a recent state-of-the-art (Chen et al. 2019), whose attribution shows signs of breakdown from the perturbation.

Related Work

We divide our discussion of related work into subsections that capture earlier efforts that are related to ours from different perspectives.

Adversarial Robustness: The possibility of fooling neural networks by crafting visually imperceptible images was first shown by (Szegedy et al. 2013). Since then, we have seen extensive efforts over the last few years in the same direction. (Goodfellow, Shlens, and Szegedy 2015) introduced one-step Fast Gradient Sign Method (FGSM) attack which was followed by more effective iterative attacks such as (Kurakin, Goodfellow, and Bengio 2016), PGD attack (Madry et al. 2018), Carlini Wagner attack (Carlini and Wagner 2017), Momentum Iterative attack (Dong et al. 2018), Diverse Input Iterative attack (Xie et al. 2019b), Jacobian-based saliency map approach (Papernot et al. 2016), etc. A parallel line of work has also emerged on finding strategies to defend against stronger adversarial attacks such as Adversarial Training (Madry et al. 2018), Adversarial Logit Pairing (Kannan, Kurakin, and Goodfellow 2018), Ensemble Adversarial Training (Tramèr et al. 2018), Parsevals Network (Cisse et al. 2017), Feature Denoising Training (Xie et al. 2019a), Latent Adversarial Training (Kumari et al. 2019), Jacobian Adversarial Regularizer (Chan et al. 2020), Smoothed Inference (Nemcovsky et al. 2019), etc. The recent work of (Zhang et al. 2019) explored the trade-off between adversarial robustness and accuracy.

Interpretability Methods: The space of work on robust attributions is based on generating neural network attributions which is itself an active area of research. These methods have an objective to compute the importance of input features based on the prediction function's output. Recent efforts in this direction include gradient-based methods (Simonyan, Vedaldi, and Zisserman 2013; Shrikumar et al. 2016; Sundararajan, Taly, and Yan 2017), propagation-based techniques (Bach et al. 2015; Shrikumar, Greenside, and Kundaje 2017; Zhang et al. 2016; Nam et al. 2019) or perturbation-based methods (Zeiler and Fergus 2014; Petsiuk, Das, and Saenko 2018; Ribeiro, Singh, and Guestrin 2016). Another recent work (Dombrowski et al. 2019) developed a smoothed explanation method that can resist manipulations, while our work aims to develop a training method (not explanation method) that is resistant to attributional attacks. Our work is based on integrated gradients (Sundararajan, Taly, and Yan 2017) which has often been used as a benchmark method (Chen et al. 2019; Singh et al. 2019) and is theoretically well-founded on axioms of attribution, and shown empirically strong performance.

Attributional Attacks and Robustness: The perspective that neural network interpretations can be broken by modifying the saliency map significantly with imperceptible input perturbations - while preserving the classifier's prediction - was first investigated recently in (Ghorbani, Abid, and Zou 2019). While the area of adversarial robustness is well explored, little progress has been made on attributional robustness i.e. finding models with robust explanations. (Chen et al. 2019) recently proposed a training methodology which achieves current state-of-the-art attributional robustness results. It showed that attributional robustness of a model can be improved by minimizing the change in input attribution w.r.t an imperceptible change in input. Our work is closest to this work, where an equal attribution change on two different input pixels are treated equally, irrespective of the original pixel attribution. This is not ideal as a pixel with high initial attribution may need to be preserved more carefully than

a pixel with low attribution. (Chen et al. 2019) has another drawback as it only considers input pixel attribution w.r.t. true class, but doesn't inspect the effect w.r.t. other negative classes. We find this to be important, and have addressed both these issues in this work. Another recent method (Singh et al. 2019) tried to achieve better attributional robustness by encouraging the observation that a meaningful saliency map of an image should be perceptually similar to the image itself. This method fails specifically for images where the true class contains objects darker than the rest of the image, or there are bright pixels anywhere in the image outside the object of interest. In both these cases, enforcing similarity of saliency to the original image objective shifts the attribution away from the true class in this method. We compare our method against both these methods in our experiments.

Background and Preliminaries

Our proposed training method requires computation of input pixel attribution through the Integrated Gradient (IG) method (Sundararajan, Taly, and Yan 2017), which has been used for consistency and fair comparison to the other closely related methods (Chen et al. 2019; Singh et al. 2019). It functions as a technique to provide axiomatic attribution to different input features proportional to their influence on the output. Computation of IG is mathematically approximated by constructing a sequence of images interpolating from a baseline to the actual image and then averaging the gradients of neural network output across these images, as shown below:

$$IG_i^f(\mathbf{x}_0, \mathbf{x}) = (\mathbf{x}^i - \mathbf{x}_0^i) \times \sum_{k=1}^{m} \frac{\partial f(\mathbf{x}_0^i + \frac{k}{m} \times (\mathbf{x}^i - \mathbf{x}_0^i))}{\partial \mathbf{x}^i} \times \frac{1}{m} \tag{1}$$

Here $f : R^n \rightarrow \mathcal{C}$ represents a deep network with $\mathcal{C}$ as the set of class labels, $\mathbf{x}_0$ is a baseline image with all black pixels (zero intensity value) and i is the pixel location on input image $\mathbf{x}$ for which IG is being computed.

Adversarial Attack: We evaluate the robustness of our model against two kinds of attacks, viz. Adversarial Attack and Attributional Attack, each of which is introduced herein. The goal of an adversarial attack is to find out minimum perturbation δ in the input space of $\mathbf{x}$ (i.e. input pixels for an image) that results in maximal change in classifier(f) output. In this work, to test adversarial robustness of a model, we use one of the strongest adversarial attacks, Projected Gradient Descent (PGD) (Madry et al. 2018), which is considered a benchmark for adv accuracy in other recent attributional robustness methods (Chen et al. 2019; Singh et al. 2019). PGD is an iterative variant of Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2015). PGD adversarial examples are constructed by iteratively applying FGSM and projecting the perturbed output to a valid constrained space S. PGD attack is formulated as follows:

$$\mathbf{x}^{i+1} = Proj_{\mathbf{x}+S}\left(\mathbf{x}^i + \alpha(\nabla_{\mathbf{x}} \mathcal{L}(\theta, \mathbf{x}^i, \mathbf{y}))\right) \tag{2}$$

Here, θ denotes the classifier parameters; input and output are represented as $\mathbf{x}$ and $\mathbf{y}$ respectively; and the classification loss function as $\mathcal{L}(\theta, \mathbf{x}, \mathbf{y})$. Usually, the magnitude of adversarial perturbation is constrained in a L_p-norm ball ($p \in \{0, 2, \infty\}$) to ensure that the adversarially perturbed example is perceptually similar to the original sample. Note that $\mathbf{x}^{i+1}$ denotes the perturbed sample at $(i+1)^{th}$ iteration.

Attributional Attack: The goal of an attributional attack is to devise visually imperceptible perturbations that change the interpretability of the test input maximally while preserving the predicted label. To test attributional robustness of a model, we use Iterative Feature Importance Attack (IFIA) in this work. As (Ghorbani, Abid, and Zou 2019) convincingly demonstrated, IFIA helps generate minimal perturbations that substantially change model interpretations, while keeping their predictions intact. The IFIA method is formally defined as below:

$$\begin{gathered} \arg\max_{\delta} D(I(\mathbf{x}; f), I(\mathbf{x} + \delta; f)) \\ \text{subject to: } ||\delta||_\infty \leq \epsilon \\ \text{such that: } \max\ f(\mathbf{x}; \theta) = \max\ f(\mathbf{x} + \delta; \theta) \end{gathered} \tag{3}$$

Here, $I(\mathbf{x}, f)$ is a vector of attribution scores over all input pixels when an input image $\mathbf{x}$ is presented to a classifier network f parameterized by θ. $D(I(\mathbf{x}; f), I(\mathbf{x}+\delta; f))$ measures the dissimilarity between attribution vectors $I(\mathbf{x}; f)$ and $I(\mathbf{x} + \delta; f)$. In our work, we choose D as Kendall's correlation computed on top-k pixels as in (Ghorbani, Abid, and Zou 2019). We describe this further in the Appendix due to space constraints.

Proposed Methodology

We now discuss our proposed robust attribution training strategy in detail, which: (i) enforces restricting true class attribution as a sparse heatmap and the negative class attribution to be a uniform distribution across the entire image; (ii) enforces the pixel attribution change caused by an imperceptible perturbation of the input image to consider the actual importance of the pixel in the original image. Both these objectives are achieved through the use of a regularizer in our training objective. The standard multiclass classification setup is considered where input-label pairs $(\mathbf{x}, \mathbf{y})$ are sampled from training data distribution $\mathcal{D}$ with a neural network classifier f, parametrized by θ. Our goal is to learn θ that provides better attributional robustness to the network.

Considering Negative Classes: We observe that "good" attributions generated for a true class form a localized (and sparse, considering the number of pixels in the full image) heatmap around a given object in the image (assuming an image classification problem setting). On the other hand, this implies that we'd like the most confusing/uncertain class attribution to not be localized, viz. i.e. resemble a uniform distribution across pixels in an image. As stated earlier, this hypothesis is inspired by the rigidity (and non-amorphous nature) of objects in images. To this end, we define the *Maximum Entropy Attributional Distribution* as a discrete uniform distribution in input pixel space as P_{ME}, where attribution score of each input pixel is equal to $\frac{1}{number\ of\ pixels}$. We also define a *True Class Attributional Distribution (P_{TCA})* as a distribution of attributions over input pixels for the true class output, denoted by f^{TCI}, when provided the perturbed image as input. Note that attributions are implemented using the IG method (as described in Sec), and P_{TCA} hence averages the gradients of the classifier's true class output when

input is varied from $\mathbf{x}_0$ to $\mathbf{x}'$. We also note that IG is simply a better estimate of the gradient, and hence can be computed w.r.t. every output class (we compute it for the true class here). Here $\mathbf{x}_0$ is a baseline reference image with all zero pixels, and $\mathbf{x}'$ represents the perturbed image. $\mathbf{x}'$ is chosen randomly within an l_∞-norm ϵ-ball around a given input $\mathbf{x}$. We represent P_{TCA} then as:

$$P_{TCA}(\mathbf{x}) = softmax\big(IG_{\mathbf{x}}^{f_{TCI}}(\mathbf{x}_0, \mathbf{x}')\big) \tag{4}$$

where $IG^{f_{TCI}}(\cdot,\cdot)$ is computed for every pixel in $\mathbf{x}$, and the softmax is applied over all P pixels in $\mathbf{x}$, i.e. $softmax(\mathbf{u}_i) = \frac{\exp(\mathbf{u}_i)}{\sum_{j\in P}\exp(\mathbf{u}_j)}$.

In a similar fashion, we define a *Negative Class Attributional Distribution (P_{NCA})*, where IG is computed for the most confusing negative class (i.e. class label with second highest probability) in a multi-class setting, or simply the negative class in a binary setting. P_{NCA} is given by:

$$P_{NCA} = softmax(IG_{\mathbf{x}}^{f_{NCI}}(\mathbf{x}_0, \mathbf{x}')) \tag{5}$$

We now define our *Class Attribution-based Contrastive Regularizer (CACR)* as:

$$\mathcal{L}_{CACR} = KL(P_{ME}||P_{NCA}) - KL(P_{ME}||P_{TCA}) \tag{6}$$

where KL stands for KL-divergence. We show how CACR is integrated into the overall loss function to minimize, later in this section. CACR enforces a skewness in the attribution map, corresponding to the true class, across an input image through the "$-KL(P_{ME}||P_{TCA})$" term, and a uniform attribution map corresponding to the most confusing negative class through the "$KL(P_{ME}||P_{NCA})$" term. The skewness in case of the true class forces the learner to focus on a few pixels in the image. This regularizer induces a contrastive learning on the training process, which is favorable to attributional robustness, as we show in our results.

Enforcing Attribution Bounds: If a pixel has a positive (or negative) attribution towards true class prediction, it may be acceptable if a perturbation makes the attribution more positive (or more negative, respectively). In other words, we would like the original pixel attribution to serve as a lower bound for a positively attributed pixel, or an upper bound for a negatively attributed pixel. If this is violated, it is likely that the attribution map may change. To implement this thought, we define $\mathcal{A}(\mathbf{x})$ as a *base attribution* i.e. computed using standard IG method attribution w.r.t the true class for the input image $\mathbf{x}$, given by:

$$\mathcal{A}(\mathbf{x}) = IG_{\mathbf{x}}^{f_{TCI}}(\mathbf{x}_0, \mathbf{x}) \tag{7}$$

Similarly, we define $\nabla\mathcal{A}(\mathbf{x})$ as the change in attribution w.r.t the true class given the perturbed image, i.e. (a similar definition is also used in (Chen et al. 2019)):

$$\nabla\mathcal{A}(\mathbf{x}) = IG_{\mathbf{x}}^{f_{TCI}}(\mathbf{x}, \mathbf{x}') = IG_{\mathbf{x}}^{f_{TCI}}(\mathbf{x}_0, \mathbf{x}') - IG_{\mathbf{x}}^{f_{TCI}}(\mathbf{x}_0, \mathbf{x}) \tag{8}$$

The abovementioned desideratum necessitates that the sign of every element of $\mathcal{A} \odot \nabla\mathcal{A}$, where $\odot$ is the element-wise/Hadamard product, be maintained positive across all pixels. To understand better, let us consider the i^{th} pixel in $\mathcal{A}$ to be positive (negative). This implies that the i^{th} pixel is positively (negatively) affecting classifier's true class prediction. In such a case, we would like the i^{th} component of $\nabla\mathcal{A}$ also to be positive (negative), i.e. it further increases the magnitude of attribution in the same direction (positive/negative, respectively) as before.

However, even when the $sign(\mathcal{A} \odot \nabla\mathcal{A})$ is positive for a pixel, we argue that an equal amount of change in attribution on a pixel with higher base attribution is more costly compared to a pixel with lower base attribution, i.e. we also would want the magnitude of each element in $\mathcal{A} \odot \nabla\mathcal{A}$ to be low, in addition to the overall sign being positive.

Our second regularizer, which we call *Weighted Attribution Change Regularizer (WACR)*, seeks to implement the above ideas. This is achieved by considering two subsets of pixels in a given input image $\mathbf{x}$: a set of pixels P_1 for which $sign(\mathcal{A} \odot \nabla\mathcal{A})$ is negative, i.e. sign($\mathcal{A}$) is not the same as sign($\nabla\mathcal{A}$); and a set of pixels P_2 for which sign($\mathcal{A}$) is same as sign($\nabla\mathcal{A}$). We then minimize the quantity below:

$$\mathcal{L}_{WACR} = S(\nabla\mathcal{A})\Big|_{p\in P_1} + S(\mathcal{A} \odot \nabla\mathcal{A})\Big|_{p\in P_2} \tag{9}$$

where we choose $S(\cdot)$ can be any size function, which we use as L_1-norm in this work, and $p \in P_1 \cup P_2$ is a pixel from the image $\mathbf{x}$. In Eqn 9, we note that:

- the first term attempts to reduce attribution change in pixels where sign($\mathcal{A}$) is not the same as sign($\nabla\mathcal{A}$). We argue that this reduction in $\nabla\mathcal{A}$ is not required for pixels in P_2, since an attribution change helps reinforce correct attributions for pixels in P_2.
- the second term attempts to lower the change in attribution more in pixels with higher base attribution. We argue that this is not required for pixels in P_1, since bringing down the attribution change irrespective of the base attribution is the focus for P_1.

Overall Optimization: We follow an adversarial training approach (Madry et al. 2018) to train the model. Adversarial training is a two-step process: an (i) Outer minimization; and an (ii) Inner maximization. The inner maximization is typically used to identify a suitable perturbation that achieves the objective of an attribution attack, and the outer minimization seeks to use the regularizers described above to counter the attack. We describe each of them below:

Outer Minimization: Our overall objective function for the outer minimization step is given by:

$$\min_\theta l_{CE}(\mathbf{x}', \mathbf{y}; \theta) + \lambda\big(\mathcal{L}_{CACR} + \mathcal{L}_{WACR}\big) \tag{10}$$

where l_{CE} is the standard cross-entropy loss used for the multi-class classification setting. We use λ as a common weighting coefficient for both regularizers, and use $\lambda = 1$ for all the experiments reported in this paper. We show effects of considering different λ values on our proposed method in in Sec . As P_{ME}, P_{NCA} and P_{TCA} are all discrete distributions, we calculate $\mathcal{L}_{CACR}$ as:

$$\sum_{i=1}^{P} P_{ME}(\mathbf{x}_i) \odot \Big(\log\frac{P_{ME}(\mathbf{x}_i)}{P_{NCA}(\mathbf{x}_i)} - \log\frac{P_{ME}(\mathbf{x}_i)}{P_{TCA}(\mathbf{x}_i)}\Big) \tag{11}$$

Method	Nat. acc.	Adv. acc.	Top-K	Kendall
Natural	**86.76%**	0.00%	8.12%	0.4978
Madry et al.	83.82%	41.91%	55.87%	0.7784
RAR	82.35%	47.06%	66.33%	0.7974
Ours	83.09%	**51.47%**	**69.50%**	**0.8121**

Table 1: Results on Flower dataset

Method	Nat. acc.	Adv. acc.	Top-K	Kendall
Natural	**90.86%**	0.01%	39.01%	0.4610
Madry et al.	85.73%	**73.01%**	46.12%	0.6251
RAR	85.44%	70.26%	72.08%	0.6747
Ours	85.45%	71.61%	**81.50%**	**0.7216**

Table 2: Results on Fashion-MNIST dataset

Method	Nat. acc.	Adv. acc.	Top-K	Kendall
Natural	**99.17%**	0.00%	46.61%	0.1758
Madry et al.	98.40%	**92.47%**	62.56%	0.2422
RAR	98.34%	88.17%	72.45%	0.3111
Ours	98.41%	89.53%	**81.00%**	**0.3494**

Table 3: Results on MNIST dataset

Method	Nat. acc.	Adv. acc.	Top-K	Kendall
Natural	**98.57%**	21.05%	54.16%	0.6790
Madry et al.	97.59%	**83.24%**	68.85%	0.7520
RAR	95.68%	77.12%	74.04%	0.7684
Ours	97.57%	82.33%	**77.15%**	**0.7889**

Table 4: Results on GTSRB dataset

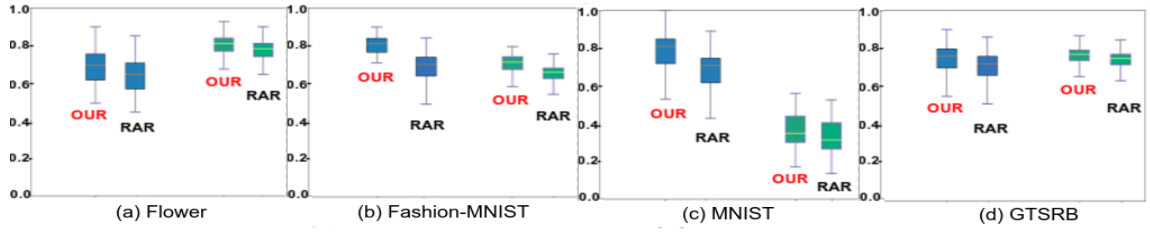

Figure 2: *(Best viewed in color)* Variance in attribution robustness metrics, Top-k intersection score and Kendall's correlation, for proposed method and RAR over all test samples on different datasets. Blue=Top-k intersection score; Green=Kendall's correlation. Our method achieves significant improvement on both metrics over RAR for all datasets, while the variance is fairly similar to variance in RAR

Method	Nat. acc.	Adv. acc.	Top-K	Kendall
Natural	**93.91%**	0.00%	38.22%	0.5643
Madry et al.	92.64%	69.85%	80.84%	0.8414
Singh et al.	93.21%	33.08%	79.84%	0.8487
Ours	90.31%	**74.26%**	**95.50%**	**0.9770**

Table 5: Results on Flower dataset (WRN 28-10 architecture)

Method	Nat. acc.	Adv. acc.	Top-K	Kendall
Natural	**99.43%**	19.9%	68.74%	0.7648
Madry et al.	98.36%	**87.49%**	86.13%	0.8842
Singh et al.	98.47%	84.66%	91.96%	0.8934
Ours	98.41%	85.17%	**92.14%**	**0.9502**

Table 6: Results on GTSRB dataset (WRN 28-10 architecture)

where P corresponds to the total number of pixels in the input image, as before. $P_{NCA}(\mathbf{x}_i)$ corresponds to the 1st order partial derivative of neural network output (corresponding to most confusing negative class) w.r.t the i^{th} input pixel. Similarly, $P_{TCA}(\mathbf{x}_i)$ corresponds to the 1st order partial derivative of neural network output (corresponding to true class) w.r.t the i^{th} input pixel.

Inner Maximization: In order to obtain the attributional attack, we use the following objective function:

$$\max_{\mathbf{x}' \in N(\mathbf{x},\epsilon)} l_{CE}(\mathbf{x}', \mathbf{y}; \theta) + S(\nabla \tilde{\mathcal{A}}) \tag{12}$$

$$\text{where } \nabla \tilde{\mathcal{A}} = IG^{\mathcal{L}_{TCI}}_{\mathbf{x}}(\mathbf{x}, \mathbf{x}')$$

Earlier computations of IG were computed w.r.t f_{TCI} or f_{NCI}, which were the softmax outputs of the true class and the most confusing negative class respectively. Here, we denote $\tilde{\mathcal{A}}$ to denote the computation of IG using the loss value corresponding to the true class. This is because our objective here is to maximize loss, while our objective was to maximize the true class softmax output in the outer minimization. We use $\mathcal{L}_{TCI}$ as the cross-entropy loss for the true class, and L_1-norm as $S(\cdot)$. Since the inner maximization is iterative by itself (and solved before the outer minimization), we randomly initialize each pixel of $\mathbf{x}'$ within an l_∞-norm ball of $\mathbf{x}$ and then iteratively maximize the objective function in Eqn 12. We avoid the use of $\mathcal{L}_{CACR}$ in our inner maximization, since $\mathcal{L}_{CACR}$ is expensive due to an extra IG calculation w.r.t. negative class, which can increase the cost due to the many iterations in the inner maximization loop.

We note that the proposed method is not an attribution method, but a training methodology that uses IG. When a model is trained using our method, all axioms of attribution will hold for IG by default, as for any other trained model. We also show that our loss function can be used as a surrogate loss of the robust prediction objective proposed by (Madry et al. 2018). Please refer to Appendix for the proof. An algorithm for our overall methodology is also presented in the Appendix due to space constraints.

Experiments and Results

We conducted a comprehensive suite of experiments and ablation studies, which we report in this section and in Sec . We report results with our method on 4 benchmark datasets i.e. Flower (Nilsback and Zisserman 2006), Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), MNIST and GTSRB (Stallkamp et al. 2012). The Flower dataset (Nilsback and Zisserman 2006) contains 17 categories with each category consisting of 40 to 258 high-definition 128×128 RGB

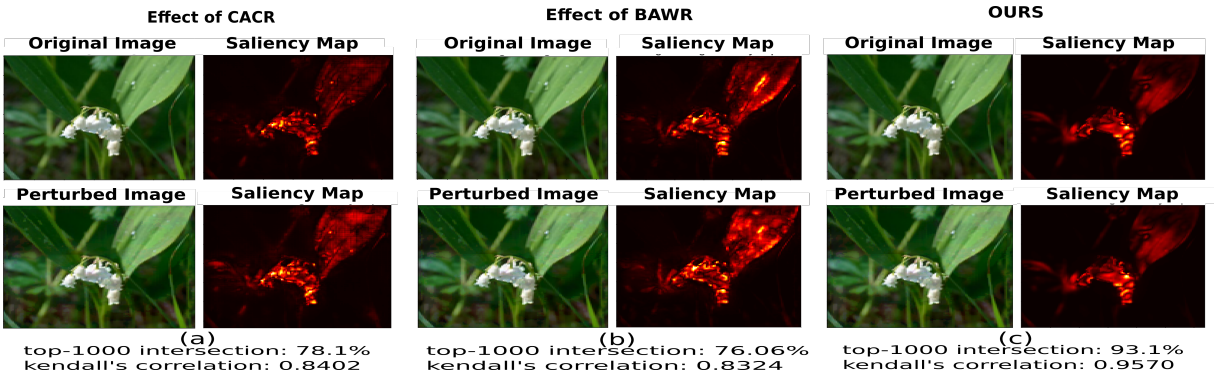

Figure 3: Visualizing effects of different regularization terms. Columns (a), (b), (c) represent saliency maps of original and perturbed image considering only $\mathcal{L}_{CACR}$, only $\mathcal{L}_{WACR}$ and both regularizers respectively in outer minimization (Eqn 10). All models predicted correctly while perturbing the original image using this attack (as in (Ghorbani, Abid, and Zou 2019))

flower images. MNIST and Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017) consist of 28×28 grayscale images from 10 categories of handwritten digits and fashion products respectively. GTSRB (Stallkamp et al. 2012) is a physical traffic sign dataset with 43 classes and around 50,000 images in total. We compare the performance of our method against existing methods: RAR (Chen et al. 2019) for attributional robustness, Singh et al (Singh et al. 2019), and Madry et al (Madry et al. 2018) which uses only standard adversarial training. Note that (Singh et al. 2019)'s code is not publicly available, and we hence compared their results only on settings reported in their paper.

Architecture Details: We used a network consisting of two convolutional layers with 32 and 64 filters respectively, each followed by 2×2 max-pooling, and a fully connected layer with 1024 neurons, for experiments with both MNIST and Fashion-MNIST datasets. We used the Resnet model in (Zagoruyko and Komodakis 2016) to perform experiments with Flower and GTSRB datasets and performed per image standardization before feeding images to the network consisting of 5 residual units with (16, 16, 32, 64) filters each. We also compared our results with a recently proposed method (Singh et al. 2019) using WRN 28-10 (Zagoruyko and Komodakis 2016) architecture as used in their paper. More architecture details for each dataset are provided in the Appendix; on any given dataset, the architectures were the same across all methods used for fair comparison.

Performance Metrics: Following (Chen et al. 2019)(Singh et al. 2019), we used top-k intersection, Kendall's correlation and Spearman correlation metrics to evaluate model's robustness against the IFIA attributional attack (Ghorbani, Abid, and Zou 2019) (Sec). Top-k intersection measures intersection size of the k most important input features before and after the attributional attack. Kendall's and Spearman correlation compute rank correlation to compare the similarity between feature importance, before and after attack. We also report natural accuracy as well as adversarial accuracy, the latter being a metric to evaluate adversarial robustness of our model against adversarial attack, such as PGD (as described in eq.2). Here, adversarial accuracy refers to the accuracy over the adversarial examples generated from perturbations on the original test set using PGD (Eqn 2).

We used a regularizer coefficient $\lambda = 1.0$ and $m = 50$ as the number of steps used for computing IG (Eqn 1) across all experiments. Note that our adversarial and attributional attack configurations were kept fixed across ours and baseline methods. Please refer the Appendix for more details on training hyperparameters and attack configurations for specific datasets.

Results: Tables 1, 2, 3 and 4 report comparisons of natural/normal accuracy, adversarial accuracy, median value of top-k intersection measure (shown as Top-K) and median value of Kendall's correlation (shown as Kendall), as used in (Chen et al. 2019), on test sets of Flower, Fashion-MNIST, MNIST and GTSRB datasets respectively. (Note that (Singh et al. 2019) did not report results on these architectures, and we report comparisons with them separately in later tables.) Our method shows significant improvement in performance on the Top-K and Kendall metrics - the metrics for attributional robustness in particular - across these datasets. Natural and adversarial accuracies are expected to be the highest for natural training and adversarial training method, Madry et al (Madry et al. 2018) respectively, and this is reflected in the results. A visual result is presented in Fig 1. More such qualitative results are presented in the Appendix. We show the variations in top-K intersection value and Kendall's correlation over all test samples for all the aforementioned 4 datasets using our method and RAR (Chen et al. 2019) in Fig 2. Our variance is fairly similar to the variance in RAR.

Tables 5 and 6 report the performance comparison (same metrics) on the Flower and GTSRB datasets using the WRN 28-10 architecture and hyperparameter settings used in (Singh et al. 2019). Note that RAR doesn't report results with this architecture, and hence is not included. We outperform Singh et al (Singh et al. 2019) by significant amounts on the Top-K and Kendall metrics, especially on the Flower dataset. A comparison with Tables 1 and 4 makes it evident that the use of the WRN 28-10 architecture leads to significant improvement in attributional robustness.

Our results vindicate the methodology proposed in this work for the state-of-the-art results obtained for attributional robustness. Although we have additional loss terms, our empirical studies showed an increase of atmost 20-25% in training time over RAR. Note that at test time, which is perhaps

Dataset	Nat. acc.	Adv. acc.	Top-K	Kendall
Flower	83.09%	49.26%	67.95%	0.8012
F-MNIST	85.43%	71.25%	78.77%	0.6974
MNIST	98.35%	88.66%	76.62%	0.3203

Table 7: Results of using only $\mathcal{L}_{CACR}$

Dataset	Nat. acc.	Adv. acc.	Top-K	Kendall
Flower	82.35%	50.00%	68.41%	0.8065
F-MNIST	85.44%	71.32%	79.11%	0.7000
MNIST	98.37%	88.78%	77.88%	0.3217

Table 8: Results of using only $\mathcal{L}_{WACR}$

Dataset	Nat. acc.	Adv. acc.	Top-K	Kendall
Flower	82.47%	50.97%	69.04	0.8101
F-MNIST	85.45%	71.56%	80.19%	0.7138
MNIST	98.39%	89.61%	79.18%	0.3337

Table 9: Proposed method with regularizer coeff for $\mathcal{L}_{CACR}$=1 and $\mathcal{L}_{WACR}$=0.7

Dataset	Nat. acc.	Adv. acc.	Top-K	Kendall
Flower	82.44%	50.81%	68.63%	0.8087
F-MNIST	85.44%	71.43%	79.86%	0.7091
MNIST	98.43%	89.59%	78.83%	0.3283

Table 10: Proposed method with regularizer coeff for $\mathcal{L}_{CACR}$=0.7 and $\mathcal{L}_{WACR}$=1

Dataset	Madry	RAR	OURS
Flower	0.7234	0.8015	**0.9004**
F-MNIST	0.7897	0.8634	**0.9289**
MNIST	0.9826	0.9928	**0.9957**
GTSRB	0.8154	0.8714	**0.9368**

Table 11: Spearman correlation between attributions from diff methods w.r.t. attribs from a naturally trained model

more important in deployment of such models, there is no additional time overhead for our method.

Ablation Studies and Analysis

Quality of Saliency maps: It is important that attributional robustness methods do not distort the explanations significantly. One way to measure the quality of the generated explanations is through the deviation of attrib maps before and after applying our method. To judge the quality of our saliency maps, we compared the attributions generated by our method with the attributions of the original image from a naturally trained model and report the Spearman correlation in Table 11 for all datasets. The results clearly show that our saliency maps change lesser from original ones than other methods. We also conducted a human Turing test to check the goodness of the saliency maps by asking 10 human users to pick a single winning saliency map that was most true to the object in a given image among (Madry, RAR, Ours). The winning rates (in same order) were: MNIST: [30%,30%,40%]; FMNIST: [20%,40%,40%]; GTSRB: [20%,30%,50%]; Flower: [0%,30%,70%], showing that our saliency maps were truer to the image than other methods, especially on more complex datasets.

Effect of Regularizers: We analyzed the effect of each regularizer term which we introduced in the outer minimization formulation in Eqn 10. For all such studies, the inner maximization setup was kept fixed. We compared attributional and adversarial accuracies, median values of Top-K intersection and Kendall's correlation achieved with and without $\mathcal{L}_{CACR}$ and $\mathcal{L}_{WACR}$ in the outer minimization. The results reported in Tables 7 and 8 suggest that the performance deteriorated substantially by removing either $\mathcal{L}_{CACR}$ or $\mathcal{L}_{WACR}$, when compared to our original results in Tables 1,2 and 3 for Flower, Fashion-MNIST and MNIST datasets.

Fig 3 shows the same effect visually with a sample test image from the Flower dataset. $\mathcal{L}_{CACR}$ not only captures the effect of positive class, but also diminishes the effect of most confusing negative class. Absence of $\mathcal{L}_{CACR}$ may hence consider attributions towards pixels which don't belong to the positive class. We can see that removing $\mathcal{L}_{CACR}$ increased the focus on a leaf which is not the true class (flower) in Fig 3(b) as compared to Fig 3(a). $\mathcal{L}_{WACR}$ penalized a large attribution change on true class pixels (i.e. pixels with high base attribution). This can be viewed from images in Fig 3(b) where keeping $\mathcal{L}_{CACR}$ forces minimal attribution change to true class pixels, compared to pixels outside the true class. Fig 3(c) shows the result of using both regularizers which shows the best performance. More such qualitative results are also provided in the Appendix.

Effect of Regularizer Coefficients: To investigate the relative effects of each proposed regularizer term, we performed experiments with other choices of regularizer coefficients. Tables 9 and 10 show the results. Our results suggest that the performance on attributional robustness drops for both cases across all datasets, when $\mathcal{L}_{CACR}$ and $\mathcal{L}_{WACR}$ are weighted lesser (original experiments had both weights to be 1). The drop is slightly more when $\mathcal{L}_{CACR}$ is weighted lesser, although this is marginal.

Additional ablation studies, including the effect of the $KL(P_{ME}||P_{NCA})$ in Eqn 6 and the use of Base Attribution in $\mathcal{L}_{WACR}$, are included in the Appendix due to space constraints.

Conclusions

In this paper, we propose two novel regularization techniques to improve robustness of deep model explanations through axiomatic attributions of neural networks. Our experimental findings show significant improvement in attributional robustness measures and put our method ahead of existing methods for this task. Our claim is supported by quantitative and qualitative results on several benchmark datasets, followed by earlier work. Our future work includes incorporating spatial smoothing on the attribution map generated by true class, which can provide sparse and localized heatmaps. We hope our findings will inspire discovery of new attributional attacks and defenses which offers a significant pathway for new developments in trustworthy machine learning.

Acknowledgements

This work has been partly supported by the funding received from MHRD, Govt of India, and Honeywell through the UAY program (UAY/IITH005). We also acknowledge IIT-Hyderabad and JICA for provision of GPU servers for the work. We thank the anonymous reviewers for their valuable feedback that improved the presentation of this work.

References

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7).

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. IEEE.

Chan, A.; Tay, Y.; Ong, Y. S.; and Fu, J. 2020. Jacobian Adversarially Regularized Networks for Robustness. In *International Conference on Learning Representations*.

Chen, J.; Wu, X.; Rastogi, V.; Liang, Y.; and Jha, S. 2019. Robust Attribution Regularization. In *Advances in Neural Information Processing Systems*, 14300–14310.

Cisse, M.; Bojanowski, P.; Grave, E.; Dauphin, Y.; and Usunier, N. 2017. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 854–863. JMLR. org.

Dombrowski, A.-K.; Alber, M.; Anders, C.; Ackermann, M.; Müller, K.-R.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, 13589–13600.

Dong, Y.; Liao, F.; Pang, T.; Hu, X.; and Zhu, J. 2018. Discovering adversarial examples with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ghorbani, A.; Abid, A.; and Zou, J. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3681–3688.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Kannan, H.; Kurakin, A.; and Goodfellow, I. 2018. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373* .

Kumari, N.; Singh, M.; Sinha, A.; Machiraju, H.; Krishnamurthy, B.; and Balasubramanian, V. N. 2019. Harnessing the vulnerability of latent layers in adversarially trained models. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2779–2785. AAAI Press.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* .

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Nam, W.-J.; Gur, S.; Choi, J.; Wolf, L.; and Lee, S.-W. 2019. Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks. *arXiv preprint arXiv:1904.00605* .

Nemcovsky, Y.; Zheltonozhskii, E.; Baskin, C.; Chmiel, B.; Bronstein, A. M.; and Mendelson, A. 2019. Smoothed Inference for Adversarially-Trained Models. *arXiv preprint arXiv:1911.07198* .

Nilsback, M.-E.; and Zisserman, A. 2006. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 1447–1454. IEEE.

Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z. B.; and Swami, A. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, 372–387. IEEE.

Petsiuk, V.; Das, A.; and Saenko, K. 2018. Rise: Randomized input sampling for explanation of black-box models. In *BMVC*.

Qin, C.; Martens, J.; Gowal, S.; Krishnan, D.; Dvijotham, K.; Fawzi, A.; De, S.; Stanforth, R.; and Kohli, P. 2019. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, 13824–13833.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD*.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3145–3153. JMLR. org.

Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713* .

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* .

Singh, M.; Kumari, N.; Mangla, P.; Sinha, A.; Balasubramanian, V. N.; and Krishnamurthy, B. 2019. On the Benefits of Attributional Robustness. *arXiv preprint arXiv:1911.13073* .

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks* 32: 323–332.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3319–3328. JMLR. org.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* .

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2018. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2020. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* .

Xie, C.; Wu, Y.; Maaten, L. v. d.; Yuille, A. L.; and He, K. 2019a. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 501–509.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019b. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2730–2739.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* .

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573* .

Zhang, J.; Lin, Z.; Brandt, J.; Shen, X.; and Sclaroff, S. 2016. Top-down neural attention by excitation backprop. *ECCV* .