Attention, Please! Adversarial Defense via Attention Rectification and Preservation

Anonymous CVPR submission

Paper ID 2103

Abstract

This study provides a new understanding of the adversarial attack problem by examining the correlation between adversarial attack and visual attention change. In particular, we observed that: (1) images with incomplete attention regions are more vulnerable to adversarial attacks; and (2) successful adversarial attacks lead to deviated and scattered attention map. Accordingly, an attention-based adversarial defense framework is designed to simultaneously rectify the attention map for prediction and preserve the attention area between adversarial and original images. The problem of adding iteratively attacked samples is also discussed in the context of visual attention change. We hope the attention-related data analysis and defense solution in this study will shed some light on the mechanism behind the adversarial attack and also facilitate future adversarial defense/attack model design.

1. Introduction

Many standard image models are recognized to be highly vulnerable to adversarial attack, which adds small perturbation to the original samples but maliciously mislead the model prediction. Extensive studies have been conducted towards designing different adversarial attack methods to fool state-of-the-art convolutional networks [4, 6, 8, 12, 16]. Applying adversarial attack in automatic visual systems like self-driving vehicle can lead to catastrophic consequences [6]. It is thus necessary to develop effective defense methods against the potential attacks.

The attempts to develop adversarial defense solutions can be coarsely classified into three groups. (1) Denoising preprocessing, transforming the input samples before feeding into the raw model, e.g., a generative adversarial network is proposed to eliminate the potential adversarial perturbation [13]. (2) Model modification, adding more layers or sub-networks and changing the loss functions of raw model, e.g., Papernot *et al.* designed a student network for knowledge distillation [3] from raw network and reduce the sensitivity to directional perturbations [9]. (3) Adversarial training, adding adversarial samples into the training set to update the model parameters, e.g., Madry *et al.* proposed to replace all clean images with adversarial images to protect against adversary [7]. It is discussed in [1] that the former two groups of defense methods largely work by obfuscating gradients, which provide only "a false sense of security" and have been successfully attacked by circumventing the gradient calculation. Adversarial training, although simple and straightforward, does not rely on obfuscated gradients and has been proved to improve model robustness by correcting sample distribution [14]. Adversarial training is recognized as a way of regularization and updates the decision boundary around adversarial samples [17].

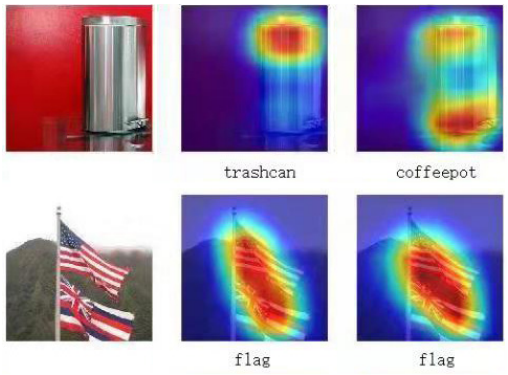


Figure 1. Visual attention of successful-attacked image (top row) and failed-attacked image (bottom row). In each row, we show the original image, original image's attention map and adversarial image's attention map. Below the attention map shows the predicted object label.

Adversarial training provides a fundamental and flexible defense framework compatible with different realizations. The performance of specific realization basically depends on the following three factors: (1) Regularizing model to focus on robust features. It is observed that adversarial perturbations contribute to amplify the importance of low-confidence features to change the output prediction [18].

Modifying models to restrict the model prediction focusing on robust features is expected to improve the robustness of original samples. (2) Reducing feature distribution divergence. In addition to injecting the adversarial samples into training set, further constraints can be designed between original and adversarial samples to alleviate feature fluctuation caused by adversarial input perturbations. (3) Selecting adversarial training samples. Tramer *et al.* claimed that the performance of adversarial training largely depends on how strong of the adversarial samples to be injected [17]. Selecting moderately strong adversarial samples therefore serves as another factor contributing to defense performance.

This study falls into the adversarial training group and attempts to address the above factors to improve the defense performance. Visual attention has been used to explain which region of image is responsible for the network's decision [19]. Through data analysis, we observed some correlations between visual attention and adversarial attack. Fig. 1 shows two example images from CIFAR-10. The attention map of original images and the corresponding adversarial images are illustrated for comparison. Quick observations include: (1) By comparing the attention maps of the two original images, we found that the upper image relies on the fractional object region for prediction and turns out vulnerable to the adversarial attack ("trashcan"→"coffeepot"). While, the lower image has a more complete and accurate region of interest and retains the predicted label. (2) By examining the change of attention map from original to adversarial images, we found that successful adversarial attack tends to deviate and scatter the attention area. The distraction of visual attention makes the prediction focusing on incomplete and wrong regions.

These attention-oriented observations inspire us to design adversarial defense solution by rectifying and preserving visual attention. The proposed Attention-based Adversarial Defense (AAD) framework consists of three components (as illustrated in Fig. 4): (1) attention rectification component, to complete and correct the prediction of original images focusing on the actual object of interest; (2) attention preservation component, to align the visual attention area between adversarial and original images to alleviate the feature divergence; and (3) adversarial training sample selection, to add moderately strong adversarial samples into training set based on the attention evolution analysis. The main contributions of this study are two-fold:

- We conducted a comprehensive data analysis and observed that successful adversarial attack exploits the incomplete attention area and brings significant fluctuation to attention map. This provides a new understanding of the adversarial attack problem from the attention perspective.
- A novel attention-based adversarial defense method is proposed to simultaneously rectify and preserve the visual attention area. Qualitative and quantitative results on MNIST and CIFAR-10 demonstrate its superior defense performance. The framework is flexible that alternative modeling of attention loss can be readily integrated into existing adversarial attack as well as defense solutions.

2. Attention-oriented Data Analysis

Visual attention helps explain to what extent each pixel of a given image contributes to the prediction of the network. Since adversarial attack is designed to change the previous prediction, we are motivated to examine the relationship between visual attention and adversarial attack. This data analysis section attempts to address the following three questions:

1. What kinds of images are vulnerable to adversarial attack?
2. How the visual attention of adversarial image deviates from the original image?
3. How visual attention changes in iterative attack and contributes to the attack result?

Before presenting data analysis setting and observations, we first make agreements on several key terms:

- *Attention map* & *attention area*: In this study, we obtain the *attention map* for a given input image x using Grad-CAM [11] [1], which is denoted as:

$$g(x) = \text{Grad-CAM}(x). \tag{1}$$

  To prevent low-contribution pixels affecting the analysis results, we further introduce *attention area* as the binary mask indicating image pixels with attention value above a threshold κ:

$$Att(x) = sign(\text{ReLU}(g(x) - \kappa)), \tag{2}$$

  where $\text{ReLU}(\cdot)$ is the activate function of network.
- *Ground-truth area*: Taking object classification as example, attention area corresponds to the region where the classification method relies to recognize certain object. This study uses *ground-truth area* to indicate the actual object region, which is obtained by object segment mask. The ground-truth area of object l in image x is denoted as $GT_l(x)$.

[1] To guarantee the derived data observations are insensitive to the choice of attention map generator, we also employed LIME [10] for data analysis and obtained consistent observations. Due to space limitation, visualization and quantitative results about LIME are provided in the supplementary material.

$\mathcal{I}_{succeed}$		$\mathcal{I}_{fail}$	
Percentage	IoU$_{Att_GT}$	Percentage	IoU$_{Att_GT}$
74.1%	0.647	25.9%	0.701

Table 1. Average IoU between attention and ground-truth area.

$\mathcal{I}_{succeed}$		$\mathcal{I}_{fail}$	
Percentage	Confidence	Percentage	Confidence
82.8%	0.8944	17.2%	0.7941

Table 2. Percentage and average confidence score of images with IoU<0.5.

- *Adversarial attack*: The adversarial examples used in data analysis, unless otherwise specified, are generated by $StepLL$ [17]:

$$x_{adv} = x_{ori} - \varepsilon \cdot sign(\bigtriangledown_x L(f(x_{ori}), y_{LL})) \quad (3)$$

where x_{ori}, x_{adv} represent original and adversarial images, ε is the step size, $f(\cdot)$ is the original network, and y_{LL} denotes the label with the lowest confidence.

2.1. Adversarial Attack Vulnerability

It is noticed that adversarial attack not always succeeds and fails on some samples. This motives us to study what characteristics make these samples robust to the attack and retain the original decision. Specifically, we examined the attention area of different images in the context of classification problem, and analyzed its correlation with the vulnerableness to adversarial attack. This is Fig.1 The data analysis was conducted with the classification network, InceptionV3 [15], and over the 50,000 images in the development set of ImageNet 2012 [2]. Since we view visual attention as support on the highest output, the 38,245 development images with the correct top-1 prediction construct the image set $\mathcal{I}_{att}$ for attention analysis.

For each image $x \in \mathcal{I}_{att}$, its attention map $g(x)$ was calculated, and the ground-truth area $GT_l(x)$ corresponding to the correctly predicted label l was also extracted. To examine whether the visual attention matches the actual object region, we made a comparison between the attention area and ground-truth area. The attention area was extracted by selecting image-specific threshold κ so that $Att(x)$ and $GT_l(x)$ have the same area size. IoU (Intersection-over-Union) between attention and ground-truth area was calculated as follows:

$$IoU_{Att\_GT}(x) = \frac{Att(x) \bigcap GT_l(x)}{Att(x) \bigcup GT_l(x)} \quad (4)$$

We separate images from $\mathcal{I}_{att}$ into two subsets, those retaining the original decision where adversarial attack failed to construct $\mathcal{I}_{fail}$, and those changing decision where adversarial attack succeeded to construct $\mathcal{I}_{succeed}$. The percentage of images falling in each subset and the corresponding average IoU_{Att_GT} are summarized in Table 1. Since all the images are correctly classified by the original network, both subsets show large IoU scores. Between the two subsets, $\mathcal{I}_{fail}$ obtains notably higher IoU than $\mathcal{I}_{succeed}$. Focusing on the 5,857 images with IoU<0.5, we examined the percentage of images falling in each subset and the average confidence score on the correct label. The results are reported in Table 2. Combining results from Table 1 and Table 2, we observed that the images with low attention IoU tend to obtain low confidence score in the targeted correct label and have higher vulnerability to be adversarially attacked.

2.2. Attention Deviation from Adversarial Attack

Adversarial samples only impose small perturbations on the original input but encounter significant change on the output prediction. This motivates us to explore where factors contribute to the non-trivial output change. This subsection studies the attention deviation from adversarial attack, and examines the consistency of the attention area between original samples and adversarial samples.

We utilized the same image set $\mathcal{I}_{att}$ for data analysis. Assuming x_{adv} represents the adversarial sample generated by attacking original sample x_{ori}, $Att(x_{ori}), Att(x_{adv})$ respectively denote the attention area of original and adversarial samples. The raw attention map generated by Grad-CAM constitutes an 8×8 grid. The attention area of original and adversarial samples are constructed by keeping the same number of grid cells with the highest attention score. Under a certain number of grid cells $\#cell$, the IoU of attention area between original samples and adversarial samples was calculated as follows:

$$IoU_{ori\_adv}(x) = \frac{Att(x_{adv}) \bigcap Att(x_{ori})}{Att(x_{adv}) \bigcup Att(x_{ori})} \quad (5)$$

Varying the number of remained grid cells from 5 to 30, we summarized the average IoU_{ori_adv} in Table 3 for $\mathcal{I}_{succeed}$ and $\mathcal{I}_{fail}$ respectively. We find a consistent result for different selections of grid cells: the IoU score of failed attack group is significantly higher than that of successful attack group. Heavy attention deviation of adversarial samples from original samples offers a strong indication of successful attack. Other than the decrease of overlap of attention areas, it is also evidenced from Fig. 1,2 and other samples that successful adversarial attack tends to make attention scattered [2]. A possible explanation for these observations is that successful adversarial perturbation on the input misleads the output prediction by distracting and scattering the original attention.

[2] We leave further discussion and utilization of the attention scatter observation in future study.

#cell	IoU_{ori_adv} for $\mathcal{I}_{succeed}$	IoU_{ori_adv} for $\mathcal{I}_{fail}$
5	0.393	0.574
10	0.490	0.677
20	0.609	0.766
30	0.676	0.807

Table 3. Average attention IoU between original and adversarial samples.

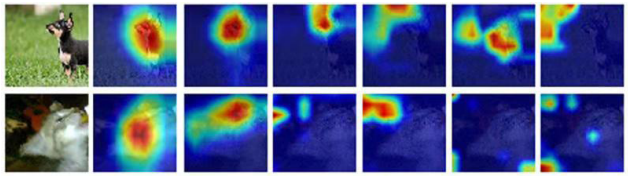


Figure 2. The change of attention map in iterative attack. (From left to right columns: original image, attention map for an original image and adversarial images after 1–5 rounds of attack.)

2.3. Attention Analysis on Iterative Attack

It is recognized that iterative adversarial attack is usually more effective than single-step attack [6]. But there is no consensus why iterative attack works better. This subsection explores the change of visual attention in iterative attack, to shed some light on this problem and inspire how to select samples for adversarial training. Fig. 2 visualizes the change of attention map in iterative attack for two images from $\mathcal{I}_{fail}$. It is shown that the attention area shrinkages and scatters during the multiple-step attacks. The two images are correctly classified as "dog" after single-step attack and misclassified as other class after the second round of attack.

In Table 4 we report the percentage of images successfully/failed attacked and their average attention IoU between original and adversarial images in each round of attack. Observations include: (1) When the adversarial perturbation is imposed in multiple rounds, the attention IoU between original and adversarial images consistently reduce. After two rounds of attack, 96.8% images are successfully attacked to change their previous prediction. This provides an attention-based explanation that iterative attack gradually deviates the original attention area and generates stronger adversarial samples. (2) When iterative attack continues (e.g., over two rounds according to data analysis), the percentage of images successfully attacked remains stable but the IoU further reduces. Lower attention IoU and scattered attention area (as shown in Fig. 2) is likely to generate "too strong" adversarial images: notable perturbations make the adversarial images visually different from the original images, which violates the intention of adversarial attack. Fig. 3 shows three examples of adversarial images after three attack rounds, which appear

Attack round	$\mathcal{I}_{succeed}$		$\mathcal{I}_{fail}$	
	Percent	IoU_{ori_adv}	Percent	IoU_{ori_adv}
1	74.1%	0.451	17.2%	0.643
2	96.8%	0.277	3.2%	0.605
3	98.1%	0.209	1.9%	0.547

Table 4. Percentage and IoU change in iterative attack.

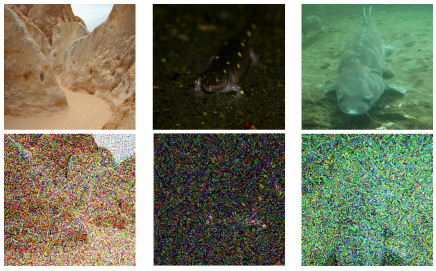


Figure 3. Example original images (top) and the corresponding adversarial images after three rounds of attack (bottom).

very differently from the original images. It is obvious that adding these "too strong" adversarial images may deteriorate the training process and lead to overfitted model.

3. Attention-based Adversarial Defense

The above data analysis demonstrates that successful adversarial perturbation leads to significant visual attention change. Our defense solution is therefore motivated to restrict the attention change to improve adversarial robustness. Specifically, observations from data analysis correspondingly inspire the three components in the proposed attention-based adversarial defense framework (as illustrated in Fig. 4): (1) attention rectification, to guide the attention area of original samples to the ground-truth area; (2) attention preservation, to punish the deviation of attention area from adversarial to original samples; and (3) adversarial training sample selection, to add moderately strong adversarial samples into the training set to prevent overfit as well as improve robustness.

3.1. Attention Rectification

As evidenced from Table 1, it is more vulnerable to adversarial attack for those samples whose prediction rely on unrelated region instead of the ground-truth area. One possible explanation is that these samples failing to focus on the actual region of interest suffer more from the adversarial perturbations and have higher risk to be misclassified. Therefore, our first component is motivated to guide the model to focus more on the ground-truth area for prediction.

Since the ground-truth area is generally unavailable during the training process, we turn to rectify the completeness of the attention area. The idea is that the attention area

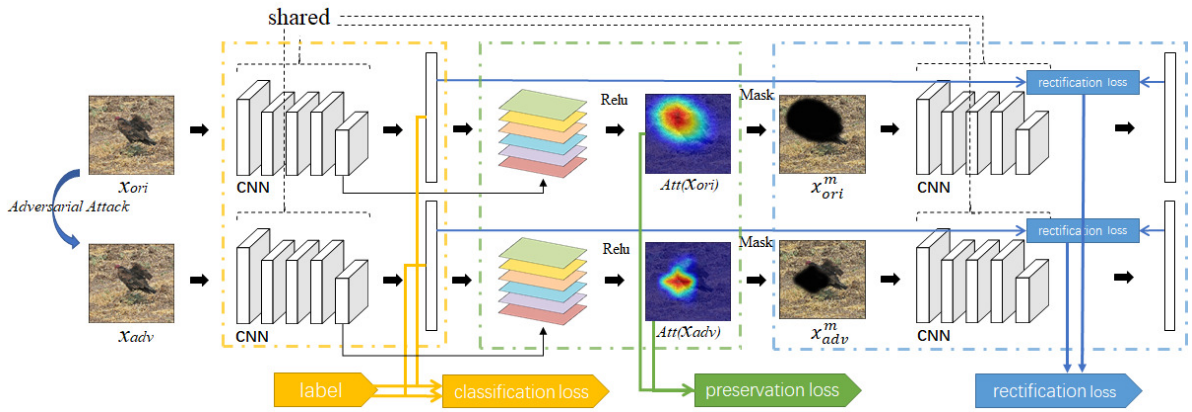


Figure 4. The proposed attention-based adversarial defense framework. The upper and lower part correspond to the training over original and adversarial samples respectively. The change of the attention map is simultaneously constrained by the rectification loss, preservation loss, and classification loss. The same parameters are shared by the four convolutional neural networks.

should include all the regions critical for prediction. In other word, the regions beyond the attention area are expected to contribute trivially to the correct prediction.

To realize this, we integrate the generation of attention area into the end-to-end training process. As illustrated on the upper part of Fig. 4, a hard mask is imposed according to the extracted attention area:

$$x^m = x \odot (1 - Att(x))), \tag{6}$$

where x is the original image, x^m denotes the image after mask, and $\odot$ denotes element-wise multiplication. To guarantee all critical regions are excluded from x^m, it is desired x and x^m lead to the prediction results as different as possible. Therefore, x and x^m are fed into the same convolutional network $f(\cdot)$ to obtain the prediction vector $f(x)$ and $f(x^m)$, and we expect the difference between $f(x)$ and $f(x^m)$ as much as possible. The same constraint is added to the adversarial image. For each original image x and the corresponding adversarial image x_{adv}, the goal is to minimize the following rectification loss:

$$L_r(x) = -\big(\mathcal{L}(f(x), f(x^m)) + \mathcal{L}(f(x_{adv}), f(x^m_{adv}))\big) \tag{7}$$

where $\mathcal{L}(\cdot,\cdot)$ denotes certain distance measure between two vectors.

3.2. Attention Preservation

It is observed from Sec. 2.2 that the shifted prediction results of adversarial image partially owes to the deviation of attention from the original image. This component attempts to preserve the attention map between original and adversarial images to reduce the influence of input perturbation to the output prediction result.

The original image x and adversarial image x_{adv} are issued to the same convolutional network to obtain attention map $g(x)$ and $g(x_{adv})$. A preservation loss is designed to minimize the pairwise difference between the two attention maps:

$$L_p(x) = \mathcal{L}\big(g(x), g(x_{adv})\big) \tag{8}$$

Combining with the rectification loss defined in the previous subsection, the overall loss for given original image x is calculated as follows:

$$L_{total}(x) = \alpha L_c(x) + \beta L_r(x) + \gamma L_p(x), \tag{9}$$

where L_c is for correct classification and defined as the standard multi-label soft margin loss, α, β and γ are weight parameters reflecting the importance of the respective component.

The three losses in Eqn. (9) jointly regularize the attention of both original image and adversarial image to the critical regions for prediction. This provides a general attention-based framework to improve adversarial robustness by rectifying and preserving visual attention. Alternative realization of the rectification loss and preservation loss are compatible with the framework.

3.3. Adversarial Training Sample Selection

The above attention rectification and preservation design constraints to make the training samples robust to potential adversarial attacks. Another critical problem remains what samples to impose these constraints to further improve the performance of adversarial training. This subsection elaborates our solution using attention map for training sample selection.

It is recognized that the defense performance of adversarial training benefits from strong attack methods to generate the adversarial samples [17]. The fact that iterative attack leads to stronger adversarial samples motivates us to also add the iteratively-attacked samples into the training

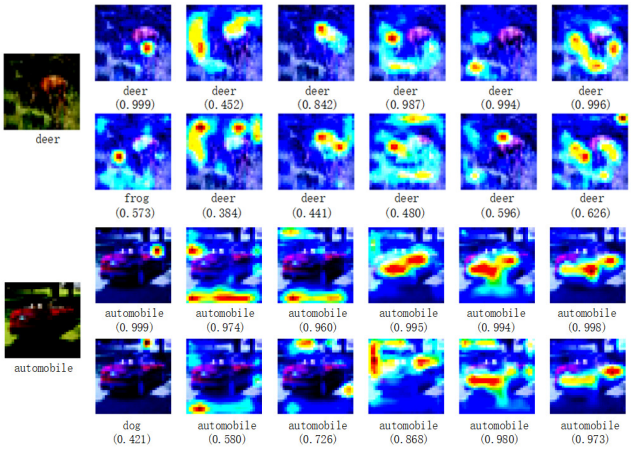


Figure 5. Attention map evolution in different training epochs. For each example image on the left, we show on the right the attention maps of its original image (top row) and adversarial image (bottom row). The six attention maps are extracted at training epoch 0, 10000, 20000, 30000, 40000 and 50000. Below each attention map shows the predicted object label and the corresponding confidence score.

set. However, as observed from Sec. 2.3, many rounds of attack potentially generates visually different images overfitting the model. A moderate round of attack is expected to satisfy the following two requirements: (1) the attack is strong enough that most adversarial images successfully change the previous prediction; and (2) the attack is not too strong to generate significantly different and even negative samples. Based on the data observations, in this study, we select a moderate round of two to generate adversarial samples and add into the training set.

Specifically, given each original image x, we iteratively generate adversarial images in two rounds as x_{adv1} and x_{adv2}. The rectification loss and preservation loss defined in Eqn.(7),(8) are modified as follows:

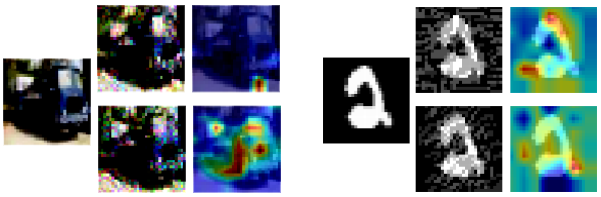
$$L_r(x) = -\big(\mathcal{L}(f(x), f(x^m)) + \mathcal{L}(f(x_{adv1}), f(x^m_{adv1})) + \mathcal{L}(f(x_{adv2}), f(x^m_{adv2}))\big),$$
$$L_p(x) = \mathcal{L}\big(g(x), g(x_{adv1})\big) + \mathcal{L}\big(g(x), g(x_{adv2})\big).$$

The influence of adding more rounds of adversarial images are provided and discussed in the experiment section.

4. Experiment

4.1. Experimental Setting

The proposed Attention-based Adversarial Defense (AAD) framework can be applied with different convolutional networks. In this study, we conducted qualitative and quantitative experiments with two simple networks, LeNet on MNIST dataset and CifarNet on CIFAR-10 dataset [5]. For Lenet, the primary parameters



(a) (b)

Figure 6. Example images from CIFAR-10 (a) and MNIST (b). On the left shows the original image. The top row on the right shows the adversarial image generated from Iter-LL with Madry's method and its corresponding attention map. The bottom row shows the results with the proposed AAD method.

are empirically set as follows: batchsize = 50, learnning rate = 0.001, iteration = 24,000, keepprop = 0.5. For CifarNet, we employ the commonly used parameters as: batchsize = 100, learnning rate = 0.0001, iteration = 50,000, keepprop = 0.5. For distance measure $\mathcal{L}(\cdot,\cdot)$ in attention losses calculation, we used l_2-norm in Eqn.(7) to consider inter-class confidences and l_1-norm in Eqn.(8) to encourage sparsity [3].

4.2. Visualization of Attention Map Evolution

The proposed AAD framework is expected to adapt the network to make predictions of both original and adversarial images focusing on the critical regions. Fig. 5 visualizes the evolution of attention map for two CIFAR-10 images as the training epoch increases.

It is shown that for each example image, on the top row, the attention area of original images is gradually rectified to the object of interest. The prediction confidence first reduces due to the consideration of attention rectification loss and preservation loss, and then recovers to guarantee high prediction confidence as well as complete attention area. On the second row, the raw attention area of adversarial image deviates much from the original image (first column on the right), which leads to the misclassification at the beginning ("deer"→"frog", "automobile"→"dog"). As the adversarial training proceeds, the attention area of adversarial image becomes consistent with that of original image, and together fit onto the object of interest at last. The confidence score for the correct object class also increases as the attention area evolves, to demonstrate the improvement of robustness against potential attacks.

4.3. Performance on Adversarial Defense

To demonstrate the defense performance of the proposed AAD framework, we compare its classification accuracy with clean model and state-of-the-art defense model from Madry *et al.* [7]. Other than StepLL, two additional

[3] Alternative distance measures are also allowed to encourage different characteristics of the attention map.

	CIFAR-10				MNIST			
	original image	StepLL	R+StepLL	Iter-LL	original image	StepLL	R+StepLL	Iter-LL
clean model	**80.6%**	2.5%	10.4%	22.1%	99.1%	46.9%	89.9%	45.2%
Madry *et al.*	78.0%	61.5%	45.4%	5.3%	**99.3%**	97.5%	98.8%	60.4%
AAD (ours)	79.1%	**74.9%**	**47.2%**	**71.5%**	99.2%	**98.1%**	**99.4%**	**61.8%**

Table 5. Defense performance comparison with Madry et al. [7] under different adversarial attack methods.

attack methods are also employed to generate adversarial samples [17]: R+StepLL, randomized single-step attack, and Iter-LL, two rounds of attack by StepLL. For CifarNet on CIFAR-10 dataset, the weight parameters α, β, γ are set in a ratio of 8:4:1. For LeNet on MNIST dataset, the weight parameters α, β, γ are set in a ratio of $2:1:2$. The obtained classification accuracy results in different settings are reported in Table 5.

On CIFAR-10 dataset, the proposed AAD method achieves superior defense performance under all the three attack methods. With the adversarial images generated from StepLL and Iter-LL, AAD obtains comparable classification accuracy with the original images. It is noted Madry's method fails to defense against Iter-LL, with classification accuracy as low as 5.3%. Fig. 6(a) shows an example image with its adversarial image and attention map from Madry's method and AAD. It is observed that Madry's method relies on the attention region beyond the object of interest for prediction. AAD manages to rectify and preserve the attention even for the strong adversarial image. Moreover, the consideration of multi-round training adversarial samples further improves AAD's robustness to iterative attacks.

On MNIST dataset, AAD achieves slightly better performance than Madry's method. It is perceived that the handwritten digit in MNIST to be classified basically positions in the center and covers a dominant region of the image. Fig. 6(b) shows such an example. In this case, the benefit from rectifying and preserving attention is limited. The proposed attention-based defense framework is more suitable for images with arbitrary object size and complex background. In the future we will validate this by evaluating AAD on ImageNet and other large-scale datasets.

The defense performance against black-box attack is also evaluated. Taking the proposed AAD defense method for example, the black-box attack is conducted as follows: (1) Two AAD models are trained under the identical configurations, denoted as *defense1* and *defense2*; (2) Adversarial images x_{adv} are generated by certain attack method (e.g., StepLL) over the model *defense1*; (3) Black-box attack is evaluated by examining the classification accuracy of x_{adv} in the other model *defense2*. Table 6 summarizes the defense performance of Madry's method and AAD against black-box attack from StepLL and Iter-LL. Consistent results are obtained with the above white-box attack:

	CIFAR-10		MNIST	
	StepLL	Iter-LL	StepLL	Iter-LL
Madry *et al.*	56.1%	35.1%	**97.1%**	86.3%
AAD (ours)	**60.5%**	**49.8%**	96.6%	**90.2%**

Table 6. Defense performance against black-box attacks.

the proposed AAD achieves superior performance than Madry's method on CIFAR-10 dataset, and comparable performance with Madry's method on MNIST dataset.

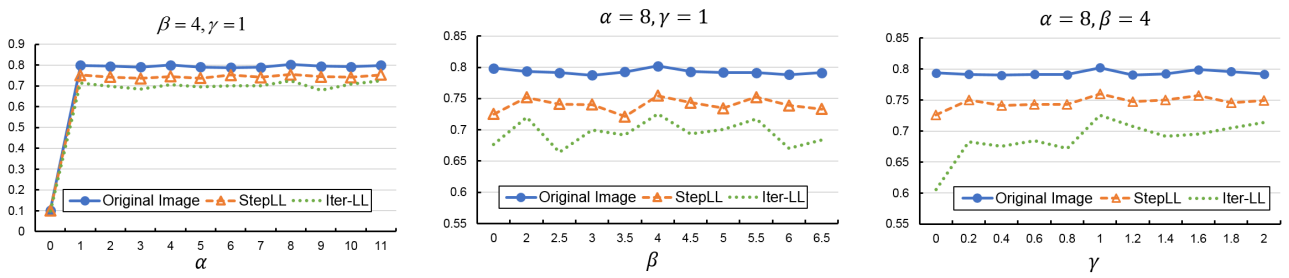
4.4. Parameter Sensitivity Analysis

The proposed attention-based defense framework mainly involves two sets of parameters: the weight parameter in Eqn. (9), and the iterative round of attack for adversarial sample generation. This subsection serves to analyze the performance sensitivity to these parameters.

We first adjusted the weight parameter to analyze the contribution of respective loss. The weight parameter sensitivity analysis experiment is conducted by fixing two of the weights and tuning the other weight. In Fig. 7(a)–(c) we show the defense classification accuracy of CIFAR-10 against different attack methods by tuning α, β, and γ respectively.

When setting the weight for respective loss as 0, i.e., excluding the corresponding classification, rectification and preservation constraint, the classification accuracy curves experience a consistently significant decline in Fig. 7(a)–(c). The sharp decrease in Fig. 7(a) is due to the fundamental role of classification loss. The notable change in Fig. 7(b)(c) validates the importance of attention rectification and preservation component towards adversarial attacks. Fig. 8 visualizes the attention map of two example images and their corresponding adversarial images w/ and w/o the proposed attention losses. The results justify our motivation to introduce the attention losses to rectify and preservation the attention area of both original and adversarial images. The best performance is generally obtained when setting $\alpha:\beta:\gamma=8:4:1$. Around this ratio the relative stable accuracy curve shows that the proposed method is not very sensitive to the weight parameter configuration within a certain range.

We further examined the influence of iterative round of attack for adversarial sample generation. The classification accuracy on CIFAR-10 is reported in Table 7 by adding 1–



(a) classification loss (b) rectification loss (c) preservation loss

Figure 7. Adversarial defense performance with different weight parameter configurations.

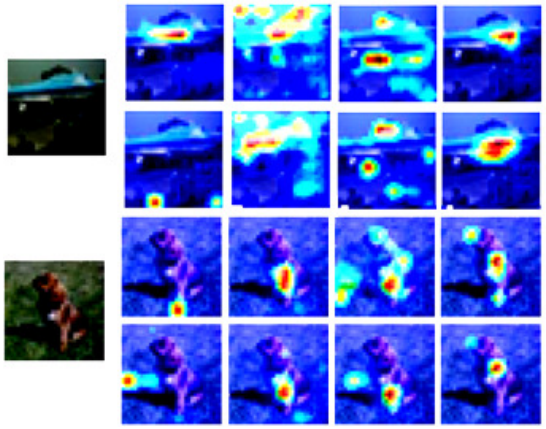


Figure 8. Attention maps from raw model, model considering rectification loss, model considering preservation loss, and model with both rectification and preservation losses.

Round	Original image	StepLL	R+StepLL	Iter-LL
1	77.7%	62.4%	44.0%	6.0%
2	79.1%	**74.9%**	**47.2%**	**71.5%**
3	**79.9%**	73.2%	38.1%	67.9%
4	79.7%	62.9%	44.1%	50.9%

Table 7. Defense classification accuracy by adding adversarial training images from different rounds of attack.

4 rounds of adversarial images into the training set. The result is basically consistent with the observation from Sec. 2.3. Considering only adversarial images from single-step attack obtains inferior defense performance especially when dealing with strong attacks like Iter-LL. Redundant rounds of attack tends to generate noisy training samples and deteriorate the training process. The rapid decline from 3 to 4 round validates this claim.

4.5. Attention-based Adversarial Attack

The data analysis observes that attention deviation contributes significantly to a successful adversarial attack. The proposed AAD method exploits this observation to improve the robustness to potential adversarial attacks. From a counter perspective, this observation also inspires the design of new attack methods by taking attention into consideration.

Original image	StepLL	StepLL+attention ($\sigma = 13$)
100.0%	26.2%	24.3%

Table 8. Adversarial attack classification accuracy on $\mathcal{I}_{att}$.

We implemented a preliminary version attention-based adversarial attack by modifying the standard StepLL in Eqn. (3). Specifically, the adversarial image is updated by finding a gradient direction amplifying the attention difference. We use l_1-norm to calculate the difference and the adversarial attack function is modified as follows:

$$x_{adv} = x_{ori} - \varepsilon \cdot sign(\nabla_x (L(f(x), y_{LL}) - \sigma \|Att(x) - g(x)\|_1)) \quad (10)$$

where σ is the weight parameter controlling the contribution of attention deviation. In this way, in addition to changing the prediction score of the most confident class, the adversarial perturbation is designed also towards deviating the attention area. Table 8 shows the classification accuracy of the 38,245 development images $\mathcal{I}_{att}$ from ImageNet2012. This preliminary result validates the effectiveness of considering attention into adversarial image generation. We emphasize that similar attention deviation term can be integrated into any gradient-based adversarial attack methods. More results and discussion are provided in the supplementary material.

5. Conclusions

This study provides a new perspective to analyze the adversarial defense/attack problem by considering attention. Qualitative and quantitative experimental results demonstrate the effectiveness of attention-based adversarial defense/attack. In the future, we are working towards seeking insight of the mechanism behind the attention perturbations from adversarial attack, as well as investigating other phenomenon concerning attention observations (e.g., the scattered attention when adversarial attack proceeds) to inspire more comprehensive defense/attack solution.

References

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018. 1

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 3

[3] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[4] C. S. Ian Goodfellow, Jonathon Shlens. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[5] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 6

[6] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 1, 4

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 6, 7

[8] D. W. Nicholas Carlini. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 1

[9] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016. 1

[10] M. T. Ribeiro, S. Singh, and C. 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. ACM, 2016. 2

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017. 2

[12] P. F. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi. Deepfool: a simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582. IEEE, 2016. 1

[13] S. Shen, G. Jin, K. Gao, and Y. Zhang. Ape-gan: Adversarial perturbation elimination with gan. *ICLR Submission, available on OpenReview*, 4, 2017. 1

[14] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017. 1

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[17] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 1, 2, 3, 5, 7

[18] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018. 1

[19] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833, 2014. 2