Adversarial Robustness via Attention Transfer

Zhuorong Li

School of Computer and Computing Science, Zhejiang University City College, Hangzhou 310015, China.

Chao Feng

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

Minghui Wu

School of Computer and Computing Science, Zhejiang University City College, Hangzhou 310015, China.

Hongchuan Yu

National Centre for Computer Animation, Bournemouth University, Poole BH12 5BB, U.K.

Jianwei Zheng

College of Computer Science and Engineering, Zhejiang University of Technology, Hangzhou 310014, China

Keywords: Adversarial defense, robustness, representation learning, visual attention, transfer learning

Abstract

Deep neural networks are known to be vulnerable to adversarial attacks. The empirical analysis in our study suggests that attacks tend to induce diverse network architectures to shift the attention to irrelevant regions. Motivated by this observation, we propose a regularization technique which enforces the attentions to be well aligned via the knowledge transfer mechanism, thereby encouraging the robustness. Our regularizer first extracts spatial attention maps that produced by the original models as additional knowledge about the training point, then feeds it into the adversarial training regimen of the defensive model. Resultant model exhibits unprecedented robustness, securing 63.81% adversarial accuracy where the prior art is 51.59% on CIFAR-10 dataset under 20-iteration PGD attacks in untargeted, white-box scenario. In addition, we go beyond performance to analytically investigate the proposed method as an effective defense. The significantly flattened loss landscape provides strong evidence that models trained by our method achieve superior performance without relying on gradient obfuscation, showing real robustness. Codes and models for our experiments are available at: https://github.com/lizhuorong/Adversarial-Robustness-via-Attention-Transfer/tree/master.

1 Introduction

Whereas deep neural networks (DNNs) have performed a broad spectrum of machine learning tasks with exceptional performance [Pouyanfar et al., 2018], they are surprisingly fragile to adversarial samples [Szegedy et al., 2013]. For example, in image classification task, by imposing imperceptible perturbation on a legitimate sample intentionally, the resultant image can drastically change the classification results [Biggio et al., 2013, Goodfellow et al., 2014a, Pei et al., 2017]. Even worse, these adversarial images can be transferred across different models[Papernot et al., 2016a, Szegedy et al., 2013], which enables the black-box adversarial attacks without any knowledge of the targeted model [Papernot et al., 2017]. It thus raises serious concerns over the robustness of these system and hinders their deployment in reliability-critical and security-sensitive applications, such as autonomous driving and identity authentication [Biggio and Roli, 2018]. It is important to note that such adversarial perturbation is an issue also common in linear classification and regression problems [Anjos and Marcel, 2011, Biggio et al., 2013, 2014, Fogla and Lee, 2006, Huang et al., 2011]. This problem has thus attracted enormous attention and encouraged high activity on adversarial defense methods [Cisse et al., 2017, Goodfellow et al., 2014b, Li et al., 2020, Szegedy et al., 2013, Tramèr et al., 2017, Yan et al., 2018], which can be roughly categorized into three catalogs: using network add-on, changing network architecture and adversarial training [Akhtar and Mian, 2018]. Our work falls into the adversarial training group and revolves around a view of model regularization.

Whereas the perturbations imposed on the original sample to craft adversarial sample are small, the change on output prediction is significant. This motivates us to shed a

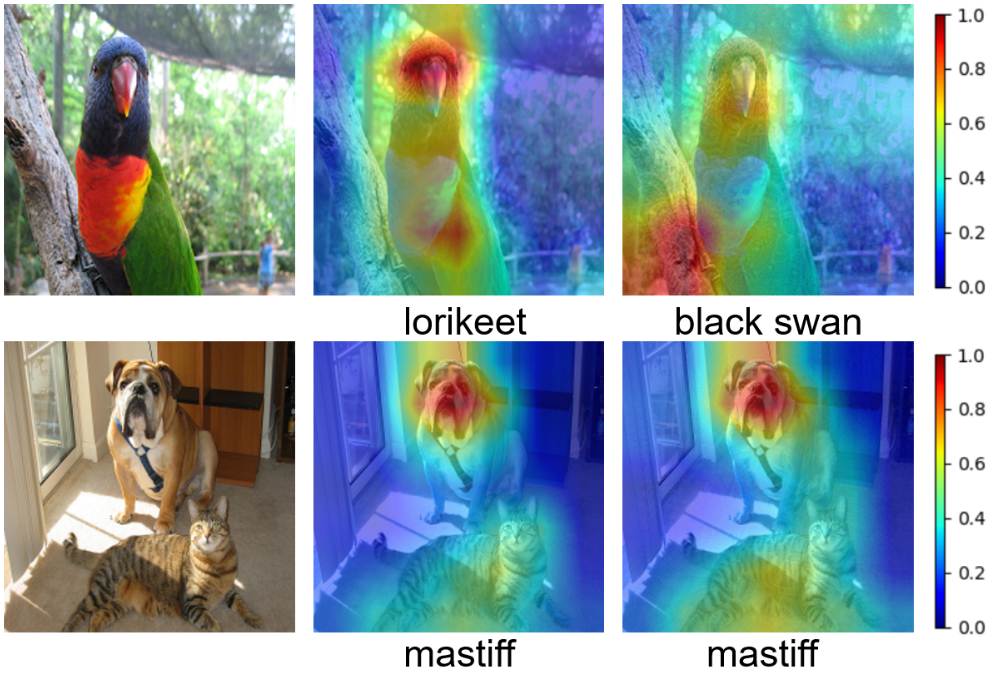

Figure 1: Correlation between attention shift and the change of prediction. From left to right columns, we show the original image, attention map for original image, and attention map for corresponding adversarial image, respectively. Adversarial images are generated w.r.t VGG16 [Simonyan and Zisserman, 2014] by PGD attack method[Madry et al., 2017]. Below the attention map we show the predicted label. Redder regions contribute more to the model prediction. It can be observed that when the attention shift is significant, the adversarial sample can successfully change the original decision of the classifier (“lorikeet”→“black swan”). Otherwise, model can still retain its prediction (“mastiff”→“mastiff”).

light on how these perturbations gradually exacerbate as the image propagates through the deep network, and finally alter the original prediction. To that end, we conduct

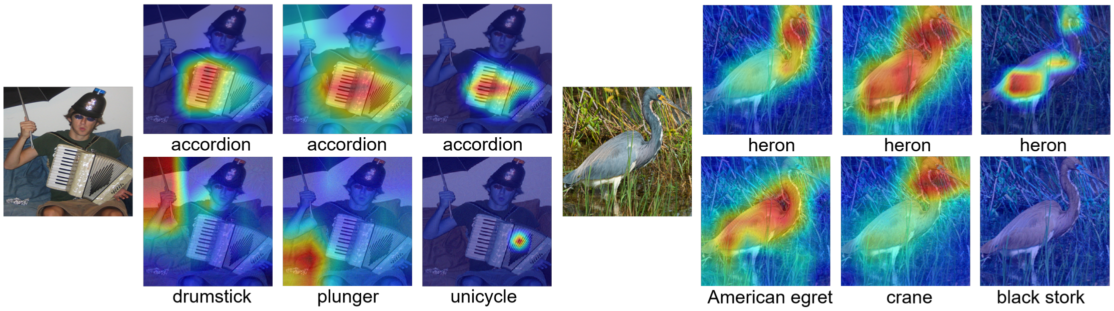

Figure 2: Examples of attention maps of different representative DNNs (VGG16 [Simonyan and Zisserman, 2014], DenseNet [Huang et al., 2017] and ResNet50 [He et al., 2016]) for prediction. For each group, we show example image on the left, and attention maps of clean image (top) and its perturbed counterpart (bottom) on the right. On the one hand, all the networks are able to focus on the class-specific object (i.e., the accordion) when given the clean images. On the other hand, attention shift or shrinkage can be widely observed when the network is applied to adversarial images, leading to erroneous prediction.

an empirical analysis of the model attention using the class activation map [Selvaraju et al., 2017, Zhou et al., 2016], which not only encodes model's interpretation regarding to the correct label of the image, but also indicates to what extent each spatial location of a given image contributes to the prediction of the network. It is thus expected to be considerably discriminative as a feature[1] and compelling as a proof-of-concept for our idea.

Through analysis, we observed close correlations between adversarial attacks and

[1] The term "feature" in this work refers to the representations of images extracted by the hidden layers of DNNs, rather than the image pixels.

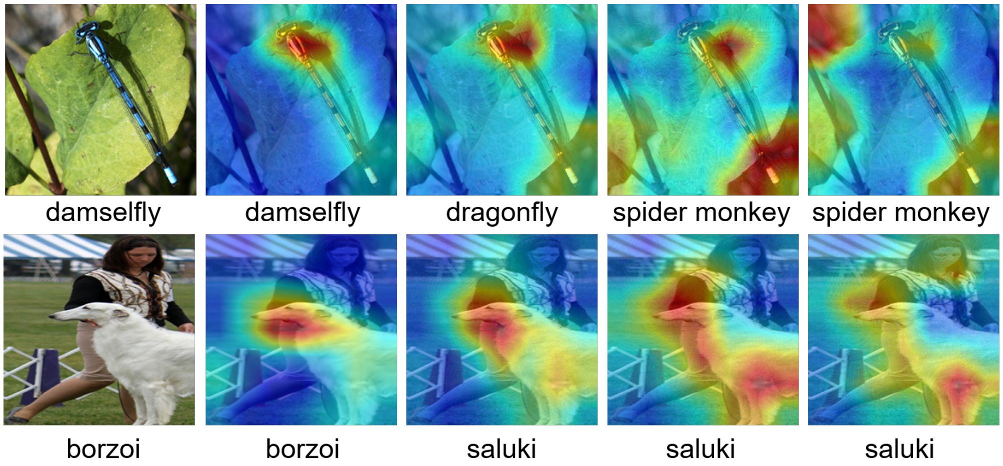

Figure 3: Evolution of the attention shift in different attack rounds under PGD attack with perturbation radius ϵ = 16/255. From left to right columns: original image and attention map for images after 0, 2, 8, 16 rounds of attack. Stronger attack generally causes more significant attention shift to change the prediction(top:“damselfly”→“spider monkey”;bottom:“borzoi”→“saluki”).

attention shifts. Specifically, Figure 1 shows that when the attention shift is significant, the adversarial sample can successfully change the original decision of the classifier (“lorikeet”→“black swan”). Otherwise, model can still retain its prediction, suggesting a causal link between the attention shift and the change of prediction. To make sure this is not specific to the network we used, we further examine across diverse network architectures. Figure 2 shows that attention areas of different architectures tend to overlap largely. For instance, all the networks focus on where the accordion lies, albeit the extensions of attention areas are not exactly the same. Besides, Figure 2 also reveals that the occurrence of attention deviation is not limited to specific networks. Moreover, we qualaitatively analyze the correlations between the strength of attacks and the deviation of attention. To that end, we use attacks optimized with different iterations to indicate attacks of different strength. As shown in Figure 3, stronger attacks generally cause more significant attention shift.

In other words, successful attacks mislead the classifier by substantially diverting or dispersing its original attention across irrelevant regions. In contrast, a classifier robust to adversarial attacks is supposed to retain its attention on object that is specific to the true class. Therefore, we are motivated to restrict the attention shifts, so that it becomes difficult for the adversary to deceive the classifier. To that end, we elaborate a regularizer on basis of the knowledge transfer, as illustrated in Figure 4. The general intuition behind the proposed regularizer is to approximate the reliable attention areas relative to true label when apply a teacher network to clean samples, and then train a student network by aligning its attention with that of the teacher, so as to suppress the fluctuation and retain attention on the class-specific object when the targeted model undergoes adversarial attacks.

Extensive evaluations show that models trained with the proposed method significantly outperform the state-of-the-art robust classifiers. They are thus suitable for deployment in security-sensitive settings. To summarize, our main contributions in this paper are threefold:

(1) We introduce an attention transfer based adversarial training (abbreviated as ATAT for simplicity), a procedure to train DNN-based models to be more resilient to adversarial attacks. It features an introduction of attention transfer mechanism, which first extracts spatial attention maps that produced by the original models as additional

knowledge about training points, then feeds it into the adversarial training regimen of the defensive model.

(2) We set new state-of-the-art adversarial accuracy on both CIFAR-10 (ϵ = 8/255) and CIFAR-100 (ϵ = 8/255) datasets under a wide range of untargeted attacks in the highly challenging white-box scenarios. Specifically, we achieve 76.74% , 63.98% , 70.15% and 63.81% adversarial accuracy on CIFAR-10 under typical attacks named FGSM, IFGSM (20), PGD (7) and PGD (20), respectively, with improvement over the previous art by up to 16%.

(3) We go beyond performance to analytically investigate the proposed method as an effective defense. We find that models trained with ATAT create flattened loss landscape, where one can move a long distance in input space without moving far in output space, thus consistent to adversarial perturbations.

2 Background and related work

We assume familiarity with neural networks [Szegedy et al., 2013], image classification, adversarial attacks [Papernot et al., 2016b] and defense [Goodfellow et al., 2014b]. We briefly review the key details and notation below.

2.1 Neural networks

Given a neural network with K layers, the feedforward dynamics induced by an input x^0 are usually denoted as

$$x^l = \begin{cases} \varphi(h^l), for \quad l = 1, ..., K-1 \\ \\ \mathrm{softmax}(h^l), \; otherwise \end{cases}$$

where φ is a non-linear function that transforms the input $h^l = W^l x^{l-1} + b^l$ into a neural activity vector x^l, W^l is a matrix of weights, and b^l is a vector of biases. We also denote the network's composite transformation in an end-to-end form and as a function $x^K = f_\theta(x^0)$, where θ denotes parameters of the network.

2.2 Image classification

To train such a network in the context of image classification, the objective is to minimize the expected risk

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(x), y)]$$

where example $x \in \mathbb{R}^d$ and the corresponding label that represented as a one-hot vector $y \in \mathbb{R}^c$ are drawn from an underlying data distribution $\mathcal{D}$, and $\mathcal{L}$ denotes the loss function, e.g., the cross-entropy loss defined by

$$\mathcal{L}_{ce}(y, f_\theta(x)) = -y^{\mathrm{T}} \mathrm{log}(f_\theta(x)).$$

Empirically, we minimize the expected risk on a finite training set $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}$ and estimate the risk on the holdout test data with average loss, which is also known as the Empirical Risk Minimization (ERM).

2.3 Adversarial attacks and adversarial training

Though models with low empirical risk work well on the holdout test set, they may degrade spectacularly in situation where adversarial examples abound due to the induced distribution shift [Biggio et al., 2013, Szegedy et al., 2013]. An intuitive defense, adversarial training, effectively alleviates the issue of distribution shift by generating the adversarial samples on-the-fly and adding them to the training set [Goodfellow et al., 2014b, Madry et al., 2017]. Essentially, it adapts the ERM paradigm to the adversarial images, towards models resistant to adversarial attacks. Many efforts have been devoted to developing adversarial training methods. First, Goodfellow *et al.* [Goodfellow et al., 2014b] propose to feed the classifier with both clean and perturbed samples generated by Fast Gradient Sigh Method (FGSM). Kurakin *et al.* [Kurakin et al., 2016] raise the issue of label leaking and suggest a replacement of FGSM that is defined w.r.t true label. However, this has been demonstrated less robust to the attacks that constructed by Tramèr *et al.* [Tramèr et al., 2017]. Madry *et al.* [Madry et al., 2017] propose a strong defense against universal attacks. Formally, they cast the adversarial training into a min-max optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max_{\delta\in\mathcal{S}_p(\epsilon)} \mathcal{L}(f_\theta(x+\delta), y)]$$

where δ is the additive perturbation that subjects to ℓ_p-norm budget ϵ. Whereas the outer optimization is to achieve resistance against adversaries by minimizing the empirical risk, the inner maximization corresponds to the generation of adversarial examples. Typically, this can be solved by gradient-based optimization, for example the one-step method such as FGSM [Goodfellow et al., 2014b].

$$x + \epsilon \cdot \mathrm{sgn} \nabla_x \mathcal{L}(f_\theta(x), y)$$

or the iterative variant, Projected Gradient Descent (PGD) method, which can be formulated as

$$x^{t+1} = \mathcal{P}_{\mathcal{S}_p(\epsilon)}(x^t + \alpha \cdot \mathrm{sgn}(\nabla_x \mathcal{L}(f_\theta(x^t), y)))$$

where sgn$(\cdot)$ is the sign function and α is the step size. We denote by $\mathcal{P}_{\mathcal{S}_p(\epsilon)}(\cdot)$ the operator that projects the input into the feasible region $\mathcal{S}_p(\epsilon)$. In particular, PGD attack is adopted in [Madry et al., 2017] to achieve universal robustness. We thus abbreviate this method to AT-PGD hereafter for simplicity. This method achieves the first empirically robust classifier on CIFAR-10 dataset and becomes the foundation of the state-of-the-art adversarial training methods.

AT-PGD provides a flexible framework compatible with various realizations, which leaves room for a flurry of activity in adversarial training. Our work uses AT-PGD as the underlying basis and revolves around a view of model regularization, to align the representation with salient data characteristics. There has been work before which focuses on the regularization techniques [Kannan et al., 2018, Mao et al., 2019, Qin et al., 2019, Zhang et al., 2019]. The work we present here is closely related to the Adversarial Logit Pairing (ALP) [Kannan et al., 2018], which highlights the similarity in logit predictions of the model for a clean sample and its adversarial counterpart, and achieves improved robustness over AT-PGD when evaluated under targeted attacks. Our method mainly differs in three aspects: the robust feature we focus on, the way how we enforce the feature alignment, and the attacks we aim to defend against. All these factors together contribute to our superiority in performance. Another prior work that inspires us is proposed by Xie *et al.* [Xie et al., 2019], which suggests architecture modifications to denoise the feature maps that hallucinated by adversarial perturbations. In contrast to it, our method does not require any architecture changes, but we are still able to denoise the contaminated feature maps.

3 Method

Armed with preliminaries briefly reviewed above, we now introduce a defense mechanism to enhance the resistance against adversarial attacks. As shown in the empirical analysis (see Figure 1-3), subtle perturbations in input space are magnified to cause substantial attention divergence between original and adversarial images, thus diverting the prediction of classifiers. Our solution is therefore motivated to retain the model attention on object that is specific to the true class. To that end, we first need to filter out attention area with back-propagated gradients provided by the true label (Section 3.1). Then we formulate the attention transfer as a regularization term to the fundamental adversarial training (Section 3.2), to align the attention of the learner on both benign and malicious samples with the benchmark attention. Finally, the attention transfer procedure is fitted into the adversarial training framework, to suit our goal of defending a DNN-based classifier against adversary (Section 3.3), as illustrated in Figure 4.

3.1 Visual Attention Extraction

Motivated to identity the attention, i.e., the hidden neuron activations w.r.t specific class, here we draw on recent work in CNN visualizations [Hendrycks and Gimpel, 2016,

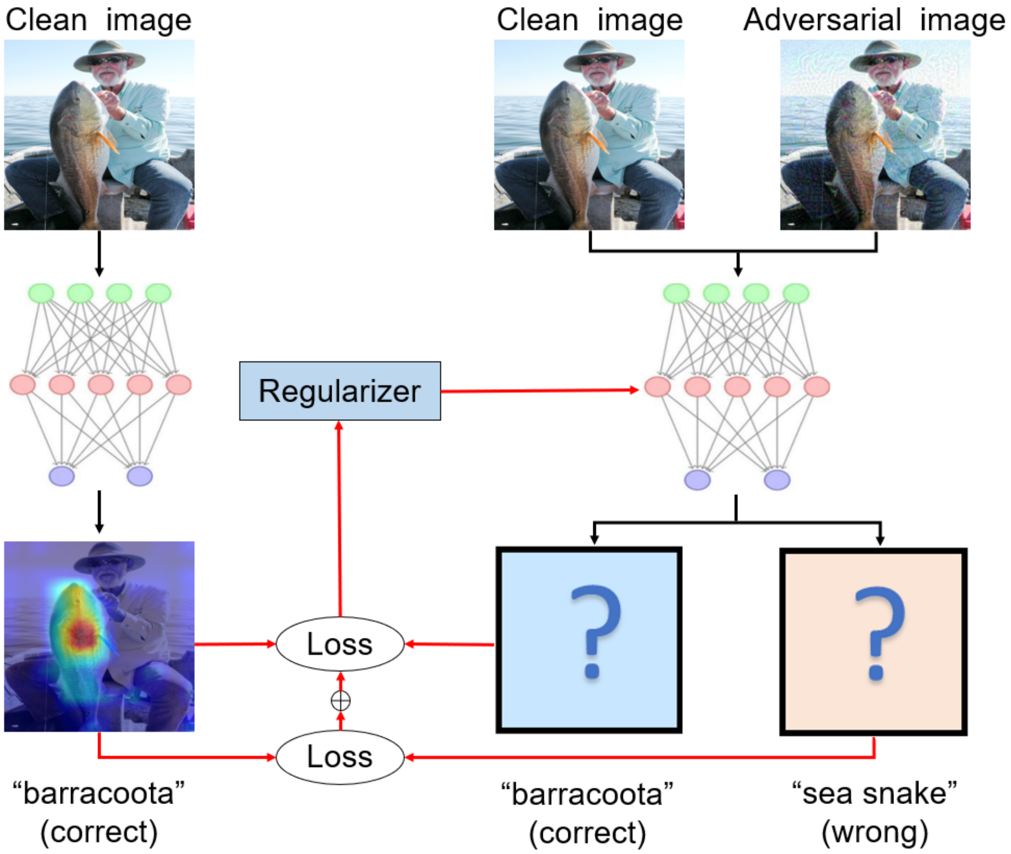

Figure 4: Schematic representation of attention transfer based regularizer. We use attention-oriented transfer learning in conjunction with the adversarial training, where the supervised signal comes from attention that activated at certain layers of the original model by clean image, and carries guidance information relative to classification to the resultant model. This model is thus expected to suppress the attention deviation that progressively amplified by small perturbation on the input, so as to make good predictions on both clean and adversarial images.

Selvaraju et al., 2017, Zhou et al., 2016]. We consider a layer of network and the corresponding activation tensor $A \in \mathbb{R}^{C \times W \times H}$, which consists of C feature maps with width W and height H. A mapping function is required to flatten this 3D tensor into a 2D spatial map as $\mathcal{M} : \mathbb{R}^{C\times W\times H} \rightarrow \mathbb{R}^{W\times H}$.

To that end, we need to compute the statistics across all the channels. With the implicit assumption that feature map in different channel can act as a basis detector for specific visual pattern, we can use a weighted combination of the presence of patterns that relevant to the final decision. Therefore, the location map can be computed as

$$L^c = \sum_k w_k^c A_k \tag{1}$$

where $A_k \in \mathbb{R}^{W\times H}$is the k-th feature map and w_k^c is the weight that corresponding to class c for k-th feature map, which can be approximated by spatially pooling the gradients:

$$w_k^c = g(\frac{\partial y^c}{\partial A_k}) \tag{2}$$

where $g(\cdot)$ denotes Global Average Pooling (GAP) as suggested in Ref [Zhou et al., 2016], to prevent overfitting as well as to preserve the location information that lost in fully-connected layers. Intuitively, the gradient-based weight above indicates to what extent A_k contributes to class c. As the weighted summation value of the spatial location is not necessarily positive, we proceed with a ReLU activation to suppress negative values that indicate confusion caused by other classes, while to only highlight features that positively impact on the attention, as Ref [Selvaraju et al., 2017]. Thus, the attention map that obtained by issuing the input x to the network Q with respect to class c is computed as:

$$ATT_Q^c(x) = \mathrm{ReLU}(\sum_k g(\frac{\partial y^c}{\partial A_k} A_k(x)) \tag{3}$$

3.2 Attention Transfer

To encourage the defensive model to focus on the ground-truth class-specific object, we use knowledge transfer [Buciluǎ et al., 2006, Hinton et al., 2015] to enable the attention of standard classifier on benign images to be transferred to the learner, i.e., the defensive network. Let us consider adversarial training in the transfer learning paradigm. Given a standardly trained classifier, we can regard it as a teacher since it provides reliable attention area that is specific to the ground-truth class when issued with clean image. Correspondingly, its defensive version plays the role of a student during training and borrows valuable class-specific knowledge from the powerful teacher model. Accordingly, the defensive network is expected to get its generalization ability improved outside the training set of the standard model, thus maintain accuracies when it encounters perturbed images. Notice that being different from the transfer learning proposed by [Hinton et al., 2015], we keep the network architecture of the student the same as that of the teacher to train, rather than simplifying the network from the complex teacher network. This is justified by our goal which is adversarial training, rather than model compression.

To encourage a student has attention maps that resembling those of the teacher, we first define transfer loss w.r.t spatial attention maps. In general, the transfer loss can be computed across multiple activation layer for which we want to proceed with attention transfer. Let θ_S and θ_T be the parameters of the student network and teacher network, respectively. $\mathcal{J}$ denotes the indices of all layers for which we want to transfer attention.

Then the attention-transfer loss can be formulated as

$$\mathcal{L}_{at} = \sum_{l \in \mathcal{J}} \| v(ATT_T^l(x_1)) - v(ATT_S^l(x_2)) \|_p \tag{4}$$

where $v(\cdot)$ denotes the vectorization operation, $v(ATT_T^l)$ and $v(ATT_S^l)$ are therefore the vectorized form of attention map that activated at the l-th layer of the teacher and the student, respectively. Notice that training data x_1 and x_2 need not to be identical according to transfer learning paradigm. Essentially, Eq.(4) can act as a metric that bridging the deviated attention to the instructive attention, which are obtained by issuing benign examples to the standard classifier.

In the adversarial training paradigm, we feed both the original image x_{ori} and its adversarial counterpart x_{adv} to the defensive network, which is initialized with the same parameters as the original network and is updated during training. Built on Eq.(4), we can align the attention by encouraging the learner to mimic the powerful teacher. Thus, the attention-transfer loss can be written as:

$$\mathcal{L}_{at} = \sum_{l \in \mathcal{J}} \| v(ATT_T^l(x_{ori})) - v(ATT_S^l(x_{adv})) \|_p.$$

$$+ \| v(ATT_T^l(x_{ori})) - v(ATT_S^l(x_{ori})) \|_p \tag{5}$$

Different norm type $\|\cdot\|_p$ can be used and we set p = 2 in our experiment to encourage inter-class confidence. As for the layer l, we choose the topmost convolution layer since higher-level visual pattern can be activated as the layer goes deeper [Bengio et al., 2013, Mahendran and Vedaldi, 2016]. However, we do not prefer the deepest layer that immediately followed with the softmax output, where critical spatial information is discarded as the attention tensor got compressed and flattened into a vector.

3.3 Attention-transfer-based Adversarial Training

With the attention-transfer loss designed above, we now can guide the attention of defensive model and accordingly alleviate the attention shifts induced by strong malicious perturbation. To fully capture the data characteristic, we also take into account the non-spatial feature, the logits vector, as suggested in [Kannan et al., 2018]:

$$\mathcal{L}_{lp} = \|p_S(x_{ori}) - p_S(x_{adv})\|_2 \tag{6}$$

where p_S is a function that maps from input image to the logits vector through the student network S. Experiment in Section 4 shows that it contributes to enhanced robustness when utilized as a supplement to our proposed loss, but not so strong when used alone. Finally, the overall loss of the proposed scheme can be calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{at} + \lambda_2 \mathcal{L}_{lp} \tag{7}$$

where $\mathcal{L}_{ce}$ is the loss function for adversarial training (i.e., the cross-entropy loss on a mixture of natural and adversarial samples for correct classification), λ_1 and λ_2 are weights that controlling the regularization effects of the respective component.

4 Experiments and Results

In this section, we first compare our method with several baselines on the CIFAR-10 and CIFAR-100 datasets. In addition, we conduct an ablation study to examine the impact of each component of our method on adversarial performance. Finally, we go beyond the performance to analyze the loss landscape and the generalization of our model.

4.1 Experimental Settings

Datasets and classifier. We empirically validate the proposed method on the popular CIFAR-10 dataset as well as the more challenging CIFAR-100 dataset [Krizhevsky et al., 2009].For both datasets, we use Wide ResNet as the classifier, which has been adopted in most of the previous works on adversarial training.

Baselines for comparison. Baseline methods we compare to include: (1) standard training using legitimate images only (Standard), (2) an unified adversarial training framework with PGD attacks (AT-PGD) [Madry et al., 2017], which is one of the most strong defenses, (3) adversarial training with additional logit pairing as a regularizer (ALP) [Huang et al., 2011] ,which is the state-of-the-art approach that has withstood intensive scrutiny, (4) a recently proposed regularization-based defense with representation learning (TLA) [Mao et al., 2019]. We use ATAT to denote the attention-transfer-based adversarial training introduced in Section 3.

Implementation details. To be comparable to the baseline methods, we follow the common protocols in [Madry et al., 2017]. Specifically, during training, we generate the adversarial samples by PGD optimization with step size of 2/255 for 7 steps. These attacks are ℓ_∞-bounded and with perturbation radius $\epsilon = 8/255$. As for testing, effective attacks including one-step attack, e.g. FGSM [Goodfellow et al., 2014a], and iteratively optimized attacks, such as I-FGSM and PGD. For PGD attacks, we run 7 and 20 iterations on CIFAR-10, and run 10 and 20 attack steps on CIFAR-100 as suggested in [Madry et al., 2017] and [Mao et al., 2019]. We consider white-box settings, where the defense is particularly challenging as the adversaries have full access to the model parameters.

Evaluation metrics. Model robustness in adversarial attack scenario is usually measured by adversarial accuracy, i.e., the percentage of times the model makes correct prediction on adversarial images under untargeted threats. In this study, we also take the nominal accuracy (i.e., the classification accuracy on natural images) into consideration to avoid sacrificing the nominal accuracy for adversarial accuracy. Besides, it is noteworthy that the adversarial accuracy we adopt in this study is a more rigorous metric than the attack success rate under targeted attacks [Engstrom et al., 2018].

4.2 Comparative Experiments

We evaluate the visual classification performance of models trained with different defense methods under white-box, untargeted attacks. As suggested in [Engstrom et al., 2018], a defense that robust against untargeted adversarial attacks is stronger than the one only robust against the targeted attacks. To close in on the true robustness, we evaluate models under a wide range of attacks. Experimental results are shown in Table 1 and 2. From these comparisons, we can observe that models trained with the proposed ATAT can comfortably outperform the others.

Experiment on CIFAR-10. We summarize the classification accuracy on the natural images (Nominal) and various malicious images in Table 1. It can be observed that Standard model completely breaks down under attacks, while AT-PGD [Madry et al., 2017], ALP [Kannan et al., 2018], TLA [Mao et al., 2019] and our ATAT separately achieves 49.70% , 52.32% 53.87%and 70.15% improvement over it under PGD(7) attacks. Notably, the proposed ATAT method achieves superior performance under all attacks with a clear margin. It is also observed that improved adversarial accuracy

Table 1: Classification accuracy comparison of the proposed method with baseline methods on CIFAR-10 dataset[1].The highest accuracy of each column is in bold to show the best performance that is specific to the attack. The proposed ATAT comfortably outperforms all the baseline methods under a wide range of attacks and sets a new state-of-the-art.

Models	Nominal	Adversial accuracy under attack(step(s))			
	accuracy	FGSM(1)	FGSM(20)	PGD(7)	PGD(20)
Standard	**95.01**	13.35	0.00	0.00	0.00
AT-PGD	87.25	56.22	45.82	49.70	45.87
ALP	86.52	60.57	46.17	52.32	46.28
TLA	86.21	58.88	51.77	53.87	51.59
ATAT	91.89	**76.74**	**63.98**	**70.15**	**63.81**

[1] Attacks crafted on CIFAR-10 are ℓ_∞-bounded with perturbation radius $\epsilon = 8/255$ under white-box setting. Baselines for comparison include AT-PGD [Madry et al., 2017], ALP [Kannan et al., 2018] and TLA [Mao et al., 2019]. We use the published results where possible.

over Standard is always accompanied by a decrease in the nominal accuracy, which is an inherent trade-off exists in all defensive models [Ilyas et al., 2019]. Nevertheless, ATAT merely gets 3.12% lower nominal accuracy than Standard, which is supposed to be outweighed by the considerable gain in adversarial robustness, e.g. 63.81% higher accuracy for the strong PGD(20)-perturbed images.

Table 2: Classification accuracy comparison of the proposed method with baseline methods on CIFAR-100 dataset[2].ATAT performs the best under a wide range of attacks, which also demonstrates that ATAT generalizes better to unseen adversarial samples.

Models	Nominal	Adversial accuracy under attack(step(s))			
	accuracy	FGSM(1)	FGSM(20)	PGD(10)	PGD(20)
Standard	**78.68**	7.8	0.00	0.00	0.00
AT-PGD	61.14	29.24	24.03	24.82	24.09
ALP	66.04	31.83	26.65	27.50	26.60
ATAT	76.22	**62.65**	**28.84**	**35.03**	**28.72**

[2] Attacks on CIFAR-100 are also ℓ_∞-bounded with perturbation radius $\epsilon = 8/255$ under white-box setting. But to the contrary of CIFAR-10 dataset, no published results of baselines on the more challenging CIFAR-100 can be found in the original paper. Therefore, we recreate all the baselines. However, TLA is not shown in the table as we are currently not able to achieve competitive results using TLA [Mao et al., 2019].

Experiment on CIFAR-100. We further turn to a more challenging dataset, CIFAR-100, to verify the proposed method. With ten times the categories of CIFAR-10 while only one-tenth images per category, CIFAR-100 greatly increases the training difficulty. Results in Table 2 hold for this challenge as well - generally lower accuracy can be observed when compared to those in Table 1. Table 2 shows the proposed ATAT stands out amongst the rest, setting a new state-of-the-art with improvement by up to 30% for single-step attack and 7% for iterative attack.

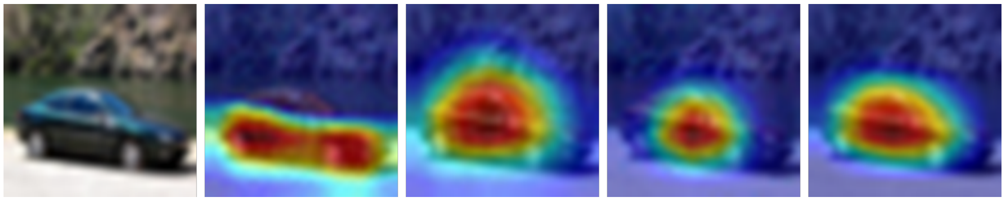

Figure 5: Ablation study by removing different components from ATAT defined in Eq.(7). From left to right: attention maps for raw model, model considering attention transfer loss, model considering logit pairing loss and full model, respectively. All the models trained with regularizer(s) show significant improvement in discrimination over the raw model, and the completed model presents the most desirable overlap with the object of interest.

The fact that undefended model is susceptible to all adversaries on both CIFAR-10 and CIFAR-100 datasets exposes the weakness of networks trained with legitimate images only, which emanates from the huge gap between classification accuracy corresponding to clean inputs and perturbed images. In contrast, models with adversarial training and examined in this study, i.e., AT-PGD, ALP, TLA and ATAT, are able to significantly narrow the gap. It is suggested that adversarial training is beneficial to enhancing model's generalization outside the training set, thus maintains accuracy when encounter perturbed image.

4.3 Ablation Study

We first compare against ablations of the full model to investigate if all components in Eq.(7) are essential. In this experiment, we use the PGD-based adversarial train-

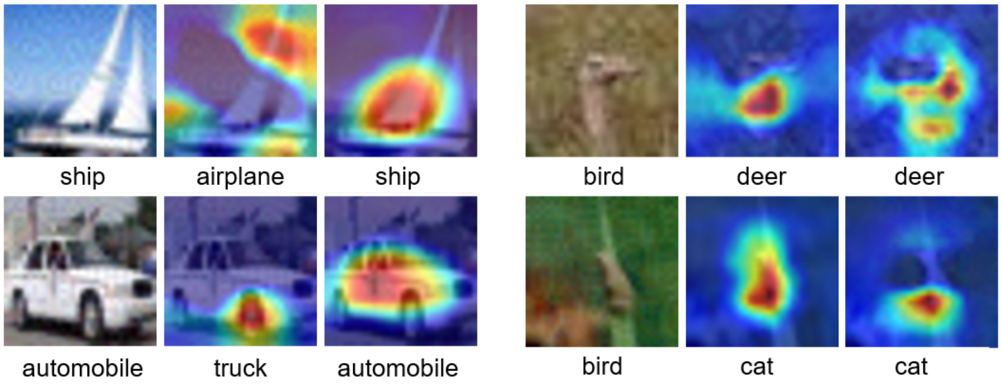

Figure 6: Effect of the proposed regularizer on rectifying model attention and thus the prediction. Left: adversarial training with the use of attention-based regularizing. Right: adversarial training without attention regularizer. For each group, we present examples of adversarial images crafted from CIFAR-10, attention maps for raw model and defended model, respectively. Attention rectified by the proposed regularizer almost perfectly overlaps the class-specific object, intuitively demonstrating the instructive effect of the proposed regularizer on attention. Subsequently, predictions can be rectified accordingly.

ing as our baseline model and activate/deactivate the component by simply setting the corresponding weight as 1/0. Figure 5 represents an example image of CIFAR-10 and attention maps from different models on the adversarially perturbed image. Since the standard model is trained on natural images only, its attention on adversarial image can barely well align with the object of interest, as shown in Figure 5. By gradually combining additional components we can observe considerable improvement. To be specific, significant boost in discrimination can be achieved by simply adding the proposed attention-based regularizer during the adversarial training, which justifies the motivation to introduce the proposed regularizer. Notice that all the models trained with additional regularizer(s) can yield considerable improvement over the raw model, suggesting that the regularization-based adversarial training is a sensible principal for adversarial defense. In these cases, model trained with the completed mode presents the most desirable overlap on the object of interest. We therefore conclude that the full mode performs the best.

To better understand the contribution of the attention-transfer-based regularizer, we further examine whether it plays a critical role in defending adversarial attacks. To do so, we visualize how the attention is rectified by adversarial training w/o the proposed regularizer and how the prediction is changed accordingly. Attention maps of raw model on adversarial samples in Figure 6, which illustrates the notable attention shift and shrinkage induced by human-undetectable adversarial perturbation, justify our motivation to enforce attention alignment. By examine the effect on attention, we found that adversarial training with our regularizer can desirably rectify the attention and thus the decision (e.g., “airplane” →“ship”, and “truck”→“automobile”). In contrast, defense solutions without explicit constraint on attention fail to align the model attention with ground-truth region of interest in many cases.

4.4 Analysis

Resistance to Gradient Obfuscation. We have already evaluated our models and demonstrated constant superiority in the comparative experiments, as presented above. In this part, we go beyond performance and analytically investigate the proposed method

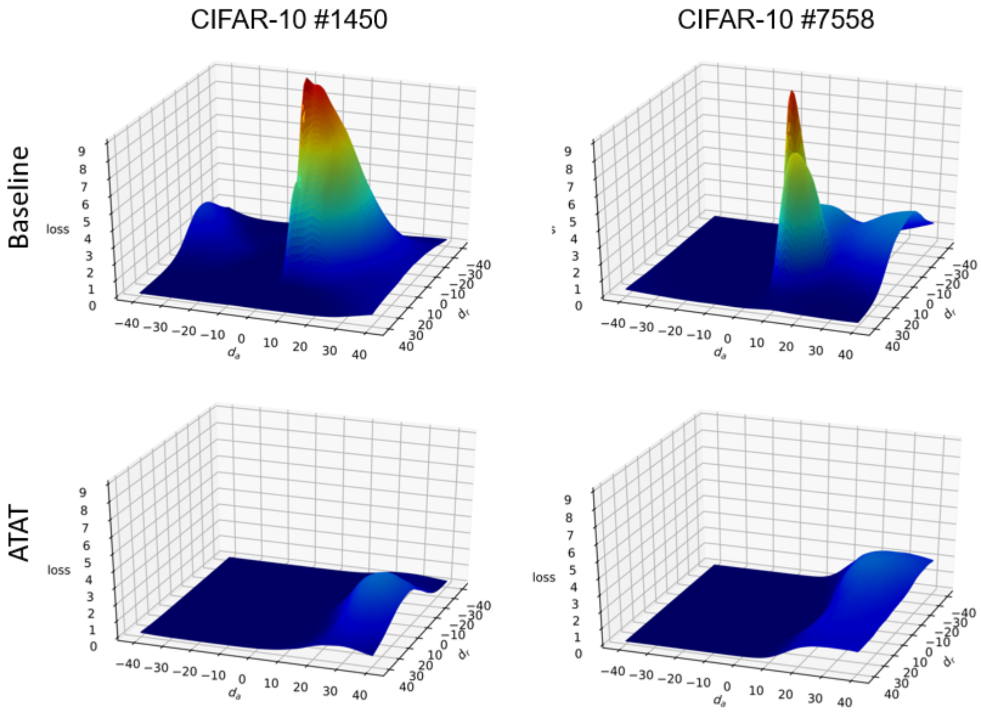

Figure 7: Resistance to gradient obfuscation. The loss landscapes before (top) and after (bottom) ATAT-based training are shown for comparison. Examples are the image 1450 (left) and 7558 (right) of CIFAR-10 test set. The substantially reduced loss and significantly flattened surface reveal real robustness achieved by our training.

through the lens of loss landscapes. This is useful for indicating whether the trained models give false sense of security, which is often due to gradient obfuscation. In particular, one type of gradient obfuscation happens when the loss landscape is highly nonlinear, which hinders the adversary from constructing an adversarial example within a few gradient-based iterations [Athalye et al., 2018, Carlini and Wagner, 2017, Uesato et al., 2018]. On the contrary, gradient obfuscation will not occur when the loss surface is flat.

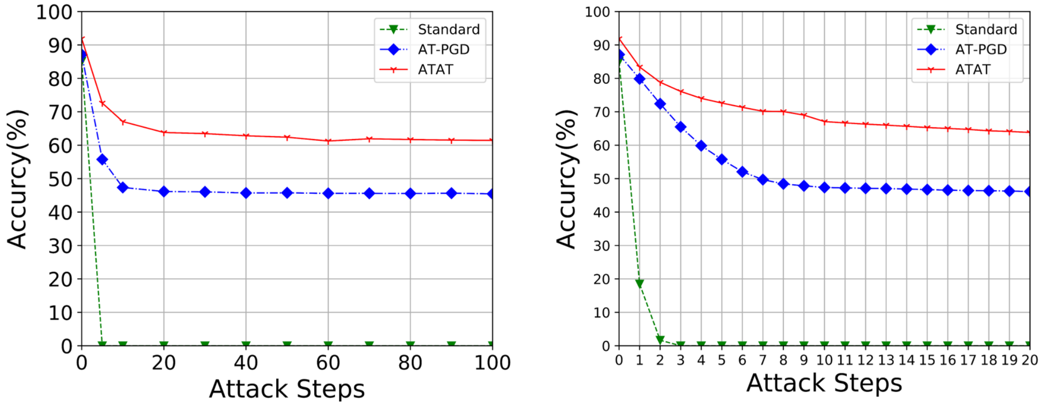

Figure 8: Classification performance of models evaluated under PGD attacks with different steps. Left: Results within 100 attack steps. Right: Zooms in the results within 20 attack steps. Defensive models (i.e., AT-PGD and ATAT) are trained on CIFAR-10 with perturbation budgets of 8/255 and steps of 7.

Figure 7 shows the comparison of loss landscapes of ATAT-trained model and raw model. Loss surface is generated by varying the input along the space defined by the adversarial perturbation ($d_a = \text{sgn}(\nabla_x f(x))$ and a random direction ($d_r = Rademacher(0.5)$), in the vicinity of the data point chosen from the test set. It can be observed that the surface of the undefended classifier (top), which is trained by standard cross-entropy over natural samples, is highly bumpy. In contrast, model trained with proposed method (bottom) not only achieves substantially reduced loss but also gives a significantly flattened surface near the input, which provides strong evidence that our superior performance is not because of gradient obfuscation, showing real robustness

Generalization to Attacks of Different Strength. We also want to see whether our defended model can well generalize to attacks of different strength. To do so, we train the defended models with attacks of 7-step optimization and evaluate the model under attacks of different iterations. The perturbation budgets are fixed at 8/255 while the PGD steps are varied from 1 to 100 to indicate attacks of different strength. Figure 8 shows our model is secured under a wide range of attack strengths. Moreover, the degradation in adversarial accuracy is more graceful for model trained with ATAT than those trained with standard adversarial training, demonstrating good generalization.

5 Conclusion

We present a regularization-based adversarial defense in this study, which draws inspiration from the evidence that adversary deceives the model by significantly distorting the high-level representation space. The proposed method distinguishes itself from others by using a transfer learning scheme for defending, which bridges the original and adversarial domain by learning the visual attention as the domain-invariant feature representation. This is beneficial to enhancing DNNs' generalization outside the training set, thus maintain accuracies when encounter adversarial images. The resultant model comfortably outperforms all the baseline methods and sets a new state-of-the-art on the CIFAR-10 dataset.

There is still room for improvement in the adversarial robustness on CIFAR-100 as we achieve the current results by simply setting the same hyper-parameters as we did on CIFAR-10. However, we do not feel such improvement necessary to show the promise of our method, as we do not strive for the end results but rather we explore how much of an effect the proposed regularizer can have on adversarial training.

We further plot the loss landscapes of the resultant model and found no obvious gradient obfuscation, suggesting its actual security. Whereas we provided explanation that matches both our intuition and the experiments results, formal analysis is still needed to get close to its real robustness. Moreover, as the anti-perturbation ability of our model relies on the representability of the training set, we plan to incorporate other attacks besides the universal gradient-based attack, such as the gradient-free attacks, hopefully advancing the robustness even further. Notice that the proposed method requires no architectural modifications and thus can enhance the robustness to adversarial attacks on most off-the-shelf DNN-based classification systems.

Acknowledgments

This work was supported in part by the EU H2020 project-AniAge under Grant 691215, the National Natural Science Foundation of China under Grant 61602413, and the Natural Science Foundation of Zhejiang Province under Grant LY19F030016.

References

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

André Anjos and Sébastien Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *2011 international joint conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Battista Biggio, Giorgio Fumera, and Fabio Roli. Pattern recognition systems under attack: Design issues and research challenges. *International Journal of Pattern Recognition and Artificial Intelligence*, 28(07):1460002, 2014.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

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

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas

Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 854–863. JMLR. org, 2017.

Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

Prahlad Fogla and Wenke Lee. Evading network anomaly detection systems: formal reasoning and practical techniques. In *Proceedings of the 13th ACM conference on Computer and communications security*, pages 59–68, 2006.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014a.

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

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.

Dan Hendrycks and Kevin Gimpel. Visible progress on adversarial images and a new saliency map. *CoRR*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136, 2019.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Zhuorong Li, Chao Feng, Jianwei Zheng, Minghui Wu, and Hongchuan Yu. Towards adversarial robustness via feature matching. *IEEE Access*, 8:88594–88603, 2020.

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.

Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3): 233–255, 2016.

Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 478–489, 2019.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram 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, 2016b.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017.

Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and SS Iyengar. A survey on deep learning:

Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5): 1–36, 2018.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, pages 13824–13833, 2019.

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

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

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.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019.

Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 419–428, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 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.