EVADING DEFENSES TO TRANSFERABLE ADVERSARIAL EXAMPLES BY MITIGATING ATTENTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks are vulnerable to adversarial examples, which can mislead classifiers by adding imperceptible perturbations. An intriguing property of adversarial examples is their good transferability, making black-box attacks feasible in real-world applications. Due to the threat of adversarial attacks, many methods have been proposed to improve the robustness, and several state-of-the-art defenses are shown to be robust against transferable adversarial examples. In this paper, we identify the attention shift phenomenon, which may hinder the transferability of adversarial examples to the defense models. It indicates that the defenses rely on different discriminative regions to make predictions compared with normally trained models. Therefore, we propose an attention-invariant attack method to generate more transferable adversarial examples. Extensive experiments on the ImageNet dataset validate the effectiveness of the proposed method. Our best attack fools eight state-of-the-art defenses at an 82% success rate on average based only on the transferability, demonstrating the insecurity of the defense techniques.

1 INTRODUCTION

Recent progress in machine learning and deep neural networks has led to substantial improvements in various pattern recognition tasks such as image understanding (Simonyan & Zisserman, 2015; He et al., 2016a), speech recognition (Graves et al., 2013), and machine translation (Sutskever et al., 2014). However, deep neural networks are highly vulnerable to adversarial examples (Biggio et al., 2013; Szegedy et al., 2014; Goodfellow et al., 2015). They are maliciously generated by adding small perturbations to legitimate examples, but make deep neural networks produce unreasonable predictions. The existence of adversarial examples, even in the physical world (Kurakin et al., 2016; Eykholt et al., 2018; Athalye et al., 2018b), has raised concerns in security-sensitive applications, e.g., self-driving cars, healthcare and finance.

Attacking deep neural networks has drawn an increasing attention since the generated adversarial examples can serve as a surrogate to evaluate the robustness of different models (Carlini & Wagner, 2017) and help to improve the robustness (Goodfellow et al., 2015; Madry et al., 2018). Several methods have been proposed to generate adversarial examples with the knowledge of the gradient information of a given model, such as fast gradient sign method (Goodfellow et al., 2015), basic iterative method (Kurakin et al., 2016), and Carlini & Wagner (2017)'s method, which are known as white-box attacks. Moreover, it is shown that adversarial examples have cross-model transferability (Liu et al., 2017), i.e., the adversarial examples crafted for one model can fool a different model with a high probability. The transferability of adversarial examples enables practical black-box attacks to real-world applications and induces serious security issues.

The threat of adversarial examples has motivated extensive research on building robust models or techniques to defend against adversarial attacks. These include training with adversarial examples (Goodfellow et al., 2015; Kurakin et al., 2017; Tramèr et al., 2018; Madry et al., 2018), image denoising/transformation (Liao et al., 2018; Xie et al., 2018a; Guo et al., 2018), leveraging generative models to move adversarial examples towards data manifold (Song et al., 2018; Samangouei et al., 2018), and theoretically-certified defenses (Raghunathan et al., 2018; Wong & Kolter, 2018). Although the non-certified defenses have demonstrated robustness against common attacks, they do so by causing obfuscated gradients, which can be easily circumvented by new attacks (Athalye et al., 2018a). However, some of the defenses (Tramèr et al., 2018; Liao et al., 2018; Xie et al., 2018a;

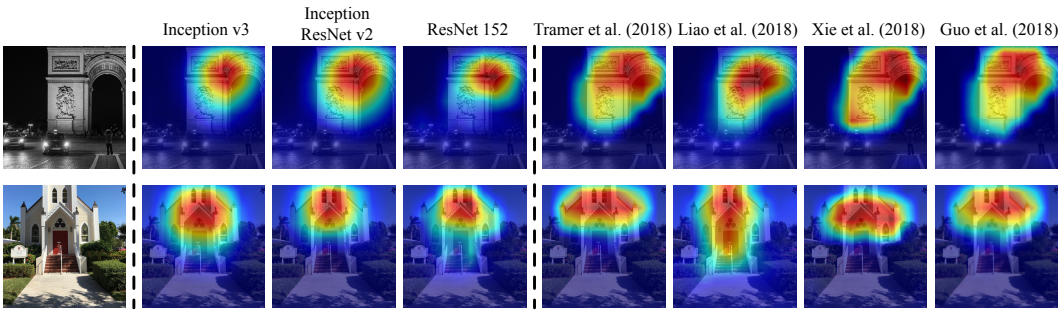


Figure 1: Demonstration of the attention shift phenomenon of the defense models compared with normally trained models. We adopt *class activation mapping* (Zhou et al., 2016) to visualize the attentive regions of three normally trained models—Inception v3 (Szegedy et al., 2016), Inception ResNet v2 (Szegedy et al., 2017), ResNet 152 (He et al., 2016a) and four defense models (Tramèr et al., 2018; Liao et al., 2018; Xie et al., 2018a; Guo et al., 2018). These defense models focus their attention on slightly different regions compared with normally trained models, which may affect the transferability of adversarial examples.

Guo et al., 2018) claim to be resistant to transferable adversarial examples, making black-box attacks difficult to evade these defenses.

In this paper, we identify *attention shift*, that the defenses make predictions based on slightly different discriminative regions compared with normally trained models, as a phenomenon which may hinder the transferability of adversarial examples to the defense models. For example, we show the attention maps of several normally trained models and defense models in Fig. 1, to represent the discriminative regions for their predictions. It is apparent that the normally trained models have similar attention maps while the defenses induce shifting attention maps. The attention shift of the defenses is caused by either training under different data distributions (Tramèr et al., 2018) or transforming the inputs before classification (Liao et al., 2018; Xie et al., 2018a; Guo et al., 2018). Therefore, the transferability of adversarial examples is largely reduced to the defenses since the structure information hidden in adversarial perturbations may be easily overlooked if a model focuses its attention on different regions.

To mitigate the effect of attention shift and evade the defenses by transferable adversarial examples, we propose an **attention-invariant** attack method. In particular, we generate an adversarial example for an ensemble of examples composed of an legitimate one and its shifted versions. Therefore the resultant adversarial example is less sensitive to the attentive region of the white-box model being attacked and may have a bigger chance to fool another black-box model with a defense mechanism based on attention shift. We further show that this method can be simply implemented by convolving the gradient with a pre-defined kernel under a mild assumption. The proposed method can be integrated into any gradient-based attack methods such as fast gradient sign method and basic iterative method. Extensive experiments demonstrate that the proposed attention-invariant attack method helps to improve the success rates of black-box attacks against the defense models by a large margin. Our best attack reaches an average success rate of 82% to evade eight state-of-the-art defenses based only on the transferability, thus demonstrating the insecurity of the current defenses.

2 Related Work

Adversarial Examples. Deep neural networks are shown to be vulnerable to adversarial examples first in the visual domain (Szegedy et al., 2014). Then several methods are proposed to generate adversarial examples for the purpose of high success rates and minimal size of perturbations (Goodfellow et al., 2015; Kurakin et al., 2016; Carlini & Wagner, 2017). They also exist in the physical world (Kurakin et al., 2016; Eykholt et al., 2018; Athalye et al., 2018b). Although adversarial examples are recently crafted for many domains, we focus on image classification tasks in this paper.

Black-box Attacks. Black-box adversaries have no access to the architecture or parameters of the target model, which are under a more challenging threat model to perform attacks. The transferability of adversarial examples provides an opportunity to attack a black-box model (Liu et al., 2017).

Several methods (Dong et al., 2018; Xie et al., 2018b) have been proposed to improve the transferability, which enable powerful black-box attacks. Besides the transfer-based black-box attacks, there is another line of works that perform attacks based on adaptive queries. For example, Papernot et al. (2017) use queries to distill the knowledge of the target model and train a surrogate model. It therefore turns the black-box attacks to the white-box attacks. Recent methods use queries to estimate the gradient or the decision boundary of the black-box model (**??**) to generate adversarial examples. However, these methods usually require tremendous number of queries, which may be impractical in real-world applications. In this paper, we resort to transferable adversarial examples for black-box attacks.

Defend against Adversarial Attacks. A large variety of methods have been proposed to increase the robustness of deep learning models. Besides directly making the models produce correct predictions for adversarial examples, some methods attempt to detect them instead (**?**Metzen et al., 2017; **?**). However most of the non-certified defenses demonstrate the robustness by causing obfuscated gradients, which are successfully circumvented by new developed attacks (Athalye et al., 2018a). Although these defenses are not robust in the white-box setting, some of them (Tramèr et al., 2018; Liao et al., 2018; Xie et al., 2018a; Guo et al., 2018) empirically show the resistance against transferable adversarial examples in the black-box setting. In this paper, we focus on generating more transferable adversarial examples against these defenses.

3 METHODOLOGY

In this section, we provide the detailed description of our algorithm. Let $\boldsymbol{x}^{real}$ denote a real example and y denote the corresponding ground-truth label. Given a classifier $f(\boldsymbol{x}) : \mathcal{X} \to \mathcal{Y}$ that outputs a label as the prediction for an input, we want to generate an adversarial example $\boldsymbol{x}^{adv}$ which is visually indistinguishable from $\boldsymbol{x}^{real}$ but fools the classifier, i.e., $f(\boldsymbol{x}^{adv}) \neq y$.[1] In most cases, the L_p norm of the adversarial perturbation is required to be smaller than a threshold ϵ as $||\boldsymbol{x}^{adv} - \boldsymbol{x}^{real}||_p \leq \epsilon$. In this paper, we use the L_∞ norm as the measurement. For adversarial example generation, the objective is to maximize the loss function $J(\boldsymbol{x}^{adv}, y)$ of the classifier, where J is often the cross-entropy loss. So the constrained optimization problem can be written as

$$\arg\max_{\boldsymbol{x}^{adv}} J(\boldsymbol{x}^{adv}, y), \quad \text{s.t. } \|\boldsymbol{x}^{adv} - \boldsymbol{x}^{real}\|_\infty \leq \epsilon. \tag{1}$$

To solve this optimization problem, the gradient of the loss function with respect to the input needs to be calculated, termed as white-box attacks. However in some cases, we cannot get access to the gradient of the classifier, where we need to perform attacks in the black-box manner. We resort to transferable adversarial examples which are generated for a different white-box classifier but have high transferability for black-box attacks.

3.1 GRADIENT-BASED ADVERSARIAL ATTACK METHODS

Several methods have been proposed to solve the optimization problem in Eq. (1). We give a brief introduction of them in this section.

Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) generates an adversarial example $\boldsymbol{x}^{adv}$ by linearizing the loss function in the input space and performing one-step update as

$$\boldsymbol{x}^{adv} = \boldsymbol{x}^{real} + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{real}, y)), \tag{2}$$

where $\nabla_{\boldsymbol{x}} J$ is the gradient of the loss function with respect to $\boldsymbol{x}$. $\text{sign}(\cdot)$ is the sign function to make the perturbation meet the L_∞ norm bound. FGSM can generate more transferable adversarial examples but is usually not effective enough for attacking white-box models (Kurakin et al., 2017).

Basic Iterative Method (BIM) (Kurakin et al., 2016) extends FGSM by iteratively applying gradient updates multiple times with a small step size α, which can be expressed as

$$\boldsymbol{x}_0^{adv} = \boldsymbol{x}^{real}, \quad \boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_t^{adv} + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)). \tag{3}$$

To restrict the generated adversarial examples within the ϵ-ball of $\boldsymbol{x}^{real}$, we can clip $\boldsymbol{x}_t^{adv}$ after each update or set $\alpha = \epsilon/T$ with T being the number of iterations. It has been shown that BIM induces

[1] This corresponds to untargeted attack. The method in this paper can be simply extended to targeted attack.

much more powerful white-box attacks than FGSM at the cost of worse transferability (Kurakin et al., 2017; Dong et al., 2018).

Momentum Iterative Fast Gradient Sign Method (MI-FGSM) (Dong et al., 2018) proposes to improve the transferability of adversarial examples by integrating a momentum term into the iterative attack method. The update procedure is

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)\|_1}, \quad \boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_t^{adv} + \alpha \cdot \text{sign}(\boldsymbol{g}_{t+1}), \tag{4}$$

where $\boldsymbol{g}_t$ gathers the gradient information up to the t-th iteration with a decay factor μ.

Diverse Inputs Iterative Fast Gradient Sign Method (Xie et al., 2018b) applies random transformations to the inputs and feeds the transformed images into the classifier for gradient calculation. The image transformation includes random resizing and padding with a given probability. This method can be combined with the momentum-based method to further improve the transferability.

Carlini & Wagner (2017)'s method is a powerful optimization-based method. It uses an auxiliary variable $\boldsymbol{v}^{adv}$ as $\boldsymbol{x}^{adv} = \frac{1}{2}(\tanh(\boldsymbol{v}^{adv}) + 1)$, and optimizes $\boldsymbol{v}^{adv}$ by solving

$$\arg\min_{\boldsymbol{v}^{adv}} \|\frac{1}{2}(\tanh(\boldsymbol{v}^{adv}) + 1) - \boldsymbol{x}^{real}\|_p - c \cdot J(\frac{1}{2}(\tanh(\boldsymbol{v}^{adv}) + 1), y), \tag{5}$$

where the loss function J could be different from the cross-entropy loss. This method aims to find adversarial examples with minimal size of perturbations, to measure the robustness of different models. It also lacks the efficacy for black-box attacks like BIM.

3.2 Attention-Invariant Attack Method

Although many attack methods (Dong et al., 2018; Xie et al., 2018b) can generate adversarial examples with very high transferability across normally trained models, they are less effective to attack defense models in the black-box manner. Some of the defenses (Tramèr et al., 2018; Liao et al., 2018; Xie et al., 2018a; Guo et al., 2018) are shown to be quite robust against black-box attacks. So we want to answer that: *Are these defenses really free from transferable adversarial examples?*

We identify the *attention shift* phenomenon which may inhibit the transferability of adversarial examples to the defenses. The attention shift refers to that the discriminative regions used by the defenses to identify object categories are slightly different from those used by normally trained models, as shown in Fig. 1. The adversarial examples generated for one model can be hardly transferred to another model with attention shift since that the structure information in adversarial perturbations may be easily destroyed if the model focuses its attention on different regions.

To reduce the effect of attention shift, we propose an **attention-invariant** attack method. In particular, rather than optimizing the objective function at a single point as Eq. (1), the proposed method uses a set of shifted images to optimize an adversarial example as

$$\arg\max_{\boldsymbol{x}^{adv}} \sum_{i,j} w_{ij} J(T_{ij}(\boldsymbol{x}^{adv}), y), \quad \text{s.t. } \|\boldsymbol{x}^{adv} - \boldsymbol{x}^{real}\|_\infty \leq \epsilon, \tag{6}$$

where $T_{ij}(\boldsymbol{x})$ is a transformation operation that shifts image $\boldsymbol{x}$ by i and j pixels along the two-dimensions respectively, i.e., each pixel (a, b) of the transformed image is $T_{ij}(\boldsymbol{x})_{a,b} = x_{a-i,b-j}$, and w_{ij} is the weight for the loss $J(T_{ij}(\boldsymbol{x}^{adv}), y)$. We set $i, j \in \{-k, ..., 0, ..., k\}$ with k being the maximal number of pixels to shift. With this method, the generated adversarial perturbations are less sensitive to the attentive regions of the white-box model, which may be transferred to another model with a higher success rate. However, we need to calculate the gradients for $(2k+1)^2$ images, which introduces much more computations. Sampling a small number of shifted images for gradient calculation is a feasible way (Athalye et al., 2018b). But we show that we can perform attacks by calculating the gradient for only one image under a mild assumption.

Convolutional neural networks are known to have the shift-invariant property (LeCun & Bengio, 1995), that an object in the input can be recognized in spite of its position. Pooling layers contribute resilience to slight transformation of the input. Therefore, we make an assumption that the shifted image $T_{ij}(\boldsymbol{x})$ is almost the same as $\boldsymbol{x}$ as inputs to the models, as well as their gradients

$$\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=T_{ij}(\hat{\boldsymbol{x}})} \approx \nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}. \tag{7}$$

Based on this assumption, we calculate the gradient of the loss defined in Eq. (6) at a point $\hat{\boldsymbol{x}}$ as

$$\begin{aligned}\nabla_{\boldsymbol{x}}\big(\sum_{i,j} w_{ij} J(T_{ij}(\boldsymbol{x}), y)\big)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} &= \sum_{i,j} w_{ij} \nabla_{\boldsymbol{x}} J(T_{ij}(\boldsymbol{x}), y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \\ &= \sum_{i,j} w_{ij} \big(\nabla_{T_{ij}(\boldsymbol{x})} J(T_{ij}(\boldsymbol{x}), y) \cdot \frac{\partial T_{ij}(\boldsymbol{x})}{\partial \boldsymbol{x}}\big)\Big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}} \\ &= \sum_{i,j} w_{ij} T_{-i-j}\big(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=T_{ij}(\hat{\boldsymbol{x}})}\big) \\ &\approx \sum_{i,j} w_{ij} T_{-i-j}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}).\end{aligned} \tag{8}$$

Given Eq. (8), we do not need to calculate the gradients for $(2k+1)^2$ images. Instead, we only need to get the gradient for the unchanged image $\hat{\boldsymbol{x}}$ and then average all the shifted gradients. This procedure is equivalent to convolving the gradient with a kernel composed of all the weights w_{ij} as

$$\sum_{i,j} w_{ij} T_{-i-j}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}) \Leftrightarrow \boldsymbol{W} * \nabla_{\boldsymbol{x}} J(\boldsymbol{x}, y)\big|_{\boldsymbol{x}=\hat{\boldsymbol{x}}}, \tag{9}$$

where $\boldsymbol{W}$ is the kernel matrix of size $(2k+1) \times (2k+1)$ with $W_{i,j} = w_{-i-j}$. In this paper, we generate the kernel $\boldsymbol{W}$ from a two-dimensional Gaussian function because: 1) the images with bigger shifts have relatively lower weights to make the adversarial perturbation fool the model at the unshifted image effectively; 2) by using a Gaussian function, this procedure is known as Gaussian blur, which is widely used in image processing.

Note that we only illustrate how to calculate the gradient of the loss function defined in Eq. (6), but do not specify the update algorithm for generating adversarial examples. This indicates that our method can be integrated into any gradient-based attack methods including FGSM, BIM, MI-FGSM, etc. Specifically, in each step we calculate the gradient $\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)$ at the current solution $\boldsymbol{x}_t^{adv}$, then convolve the gradient with the pre-defined kernel $\boldsymbol{W}$, and finally get the new solution $\boldsymbol{x}_{t+1}^{adv}$ following the update rule in different attack methods (In FGSM, there is only one step of update).

4 Experiments

In this section, we present the experimental results to demonstrate the effectiveness of the proposed method on improving the transferability of adversarial examples to the defense models.

4.1 Experimental Settings

We use an ImageNet-compatible dataset[2] comprised of 1000 images to conduct experiments. This dataset was used in the NIPS 2017 adversarial competition. We include eight defense models which are shown to be robust agsinst black-box attacks on the ImageNet dataset. These are

- Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, IncRes-v2$_{ens}$ (Tramèr et al., 2018);
- high-level representation guided denoiser (HGD, rank-1 submission in the NIPS 2017 defense competition) (Liao et al., 2018);
- input transformation through random resizing and padding (R&P, rank-2 submission in the NIPS 2017 defense competition) (Xie et al., 2018a);
- input transformation through JPEG compression or total variance minimization (TVM) (Guo et al., 2018);
- rank-3 submission[3] in the NIPS 2017 defense competition (NIPS-r3).

To attack these defenses based on the transferability, we also include four normally trained models—Inception v3 (Inc-v3) (Szegedy et al., 2016), Inception v4 (Inc-v4), Inception ResNet v2 (IncRes-v2) (Szegedy et al., 2017), and ResNet v2-152 (Res-v2-152) (He et al., 2016b), as the white-box models to generate adversarial examples.

[2] https://github.com/tensorflow/cleverhans/tree/master/examples/nips17_adversarial_competition/dataset

[3] https://github.com/anlthms/nips-2017/tree/master/mmd

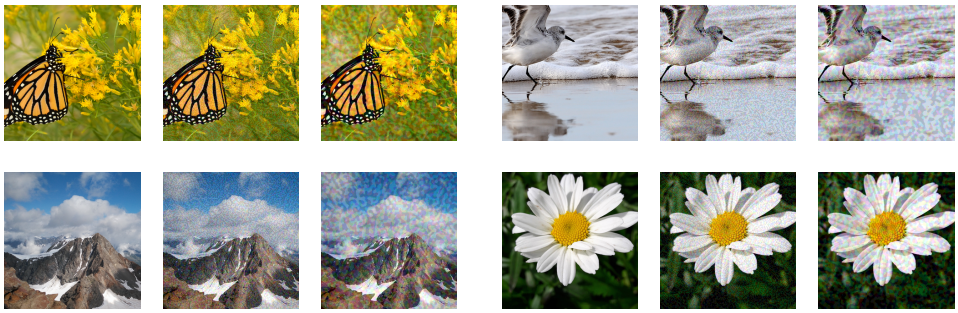


Figure 2: The adversarial examples generated for Inc-v3 using FGSM and A-FGSM.

Table 1: The success rates (%) of black-box attacks against eight defenses. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using FGSM and A-FGSM.

	Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NIPS-r3
Inc-v3	FGSM	15.6	14.7	7.0	2.1	6.5	19.9	18.8	9.8
	A-FGSM	**28.2**	**28.9**	**22.3**	**18.4**	**19.8**	**25.5**	**30.7**	**24.5**
Inc-v4	FGSM	16.2	16.1	9.0	2.6	7.9	21.8	19.9	11.5
	A-FGSM	**28.2**	**28.3**	**21.4**	**18.1**	**21.6**	**27.9**	**31.8**	**24.6**
IncRes-v2	FGSM	18.0	17.2	10.2	3.9	9.9	24.7	23.4	13.3
	A-FGSM	**32.8**	**33.6**	**28.1**	**25.4**	**28.1**	**32.4**	**38.5**	**31.4**
Res-v2-152	FGSM	20.2	17.7	9.9	3.6	8.6	24.0	22.0	12.5
	A-FGSM	**34.6**	**34.5**	**27.8**	**24.4**	**27.4**	**32.7**	**38.1**	**30.1**

In our experiments, we integrate our method into the fast gradient sign method (FGSM) (Goodfellow et al., 2015), momentum iterative fast gradient sign method (MI-FGSM) (Dong et al., 2018) and diverse input iterative fast gradient sign method with momentum (DIM) (Xie et al., 2018b). We do not include the basic iterative method and Carlini & Wagner (2017)'s method since that they are not good at generating transferable adversarial examples (Dong et al., 2018). We denote the attacks combined with our attention-invariant method as A-FGSM, A-MI-FGSM and A-DIM respectively.

For the settings of hyper-parameters, we set the maximum perturbation to be $\epsilon = 16$ among all experiments with pixel value in $[0, 255]$. For the iterative attack methods, we set the number of iteration as 10 and the step size as $\alpha = 1.6$. For MI-FGSM and A-MI-FGSM, we adopt the default dacay factor $\mu = 1.0$. For DIM and A-DIM, the transformation probability is set to 0.7. Please note that the settings for each attack method and its attention-invariant version are the same, because our method is not concerned with the specific attack precedure.

4.2 Single-Model Attacks

We first perform adversarial attacks for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using FGSM, MI-FGSM, DIM and their extensions by combining with the proposed attention-invariant attack method as A-FGSM, A-MI-FGSM and A-DIM. We then use the generated adversarial examples to attack the eight defense models we consider based only on the transferability. We report the success rates of black-box attacks in Table 1, Table 2 and Table 3, where the success rates are the misclassification rates of the corresponding defense models with adversarial images as inputs. In the attention-invariant based attacks, we set the size of the kernel matrix $\boldsymbol{W}$ as 15×15 across all experiments, and we will study the effect of kernel size in Section 4.4.

From the tables, we observe that the success rates against the defenses are improved by a large margin when using the proposed method regardless of the attack algorithms or the white-box models being attacked. In general, the attention-invariant based attacks consistently outperform the baseline attacks by $5\% \sim 30\%$. In particular, when using A-DIM, the combination of our method and DIM, to attack the IncRes-v2 model, the resultant adversarial examples have about 60% success rates against the defenses (as shown in Table 3). It demonstrates the vulnerability of the current defenses against black-box attacks. The results also validate the effectiveness of the proposed method. Although we only compare the results of our attack method with baseline methods against the defense

Table 2: The success rates (%) of black-box attacks against eight defenses. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using MI-FGSM and A-MI-FGSM.

	Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NIPS-r3
Inc-v3	MI-FGSM	20.5	17.4	9.5	6.9	8.7	20.3	19.4	12.9
	A-MI-FGSM	**35.8**	**35.1**	**25.8**	**25.7**	**23.9**	**28.2**	**34.9**	**26.7**
Inc-v4	MI-FGSM	22.1	20.1	12.1	9.6	12.1	26.0	24.8	15.6
	A-MI-FGSM	**36.7**	**39.2**	**28.7**	**27.8**	**28.0**	**31.6**	**38.4**	**29.5**
IncRes-v2	MI-FGSM	31.3	27.2	19.7	19.6	18.6	31.6	34.4	22.7
	A-MI-FGSM	**50.7**	**51.7**	**49.3**	**45.1**	**45.2**	**45.9**	**55.4**	**46.2**
Res-v2-152	MI-FGSM	25.1	23.7	13.3	15.1	14.6	31.2	24.5	18.0
	A-MI-FGSM	**39.9**	**37.7**	**32.8**	**31.8**	**31.1**	**38.3**	**41.2**	**34.4**

Table 3: The success rates (%) of black-box attacks against eight defenses. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using DIM and A-DIM.

	Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NIPS-r3
Inc-v3	DIM	24.2	24.3	13.0	9.7	13.3	30.7	24.4	18.0
	A-DIM	**46.9**	**47.1**	**37.4**	**38.3**	**36.8**	**37.0**	**44.2**	**41.4**
Inc-v4	DIM	28.3	27.5	15.6	14.6	17.2	38.6	29.1	14.1
	A-DIM	**48.6**	**47.5**	**38.7**	**40.3**	**39.3**	**43.5**	**45.6**	**41.9**
IncRes-v2	DIM	41.2	40.0	27.9	32.4	30.2	47.2	41.7	37.6
	A-DIM	**61.3**	**60.1**	**59.5**	**58.7**	**61.4**	**55.7**	**66.2**	**61.5**
Res-v2-152	DIM	40.5	36.0	24.1	32.6	26.4	42.4	36.8	34.4
	A-DIM	**56.1**	**55.5**	**49.5**	**51.8**	**50.4**	**50.8**	**55.7**	**52.9**

models, our attacks remain the success rates of baseline attacks in the white-box setting and the black-box setting against normally trained models, which will be shown in the Appendix.

We show several adversarial images generated for the Inc-v3 model by FGSM and A-FGSM in Fig. 2. It can be seen that by using A-FGSM, in which the gradients are convolved by a kernel $\boldsymbol{W}$ before applying to the raw images, the adversarial perturbations are much smoother than those generated by FGSM. The smooth effect also exists in other attention-invariant based attacks.

4.3 Ensemble-based Attacks

In this section, we further present the results when adversarial examples are generated for an ensemble of models. Liu et al. (2017) have shown that attacking multiple models at the same time can improve the transferability of the generated adversarial examples. It is due to that if an example remains adversarial for multiple models, it is more likely to transfer to another black-box model.

We adopt the ensemble method proposed by Dong et al. (2018), which fuses the logit activations of different models. We attack the ensemble of Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 with equal ensemble weights using FGSM, A-FGSM, MI-FGSM, A-FGSM, DIM and A-DIM respectively. We also set the kernel size in the attention-invariant based attacks as 15×15.

In Table 4, we show the results of black-box attacks against the eight defenses. The proposed method also improves the success rates across all experiments over the baseline attacks. It should by noted that *the adversarial examples generated by A-DIM can fool the state-of-the-art defenses at an* 82% *success rate on average based on the transferability*. And the adversarial examples are generated for normally trained models unaware of the defense strategies. The results in the paper demonstrate that the current defenses are far from real security, and cannot be deployed in real-world applications.

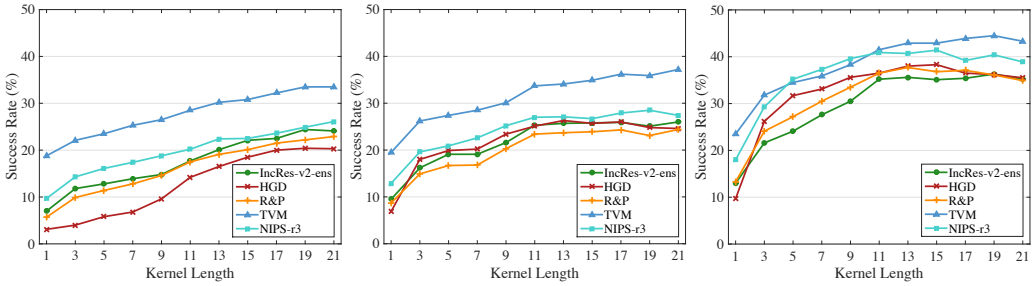
4.4 The Effect of Kernel Size

The size of the kernel $\boldsymbol{W}$ plays a key role for improving the success rates of black-box attacks. If the kernel size equals to 1×1, the attention-invariant based attacks degenerate to their vanilla versions. Therefore, we conduct an ablation study to examine the effect of different kernel sizes.

We attack the Inc-v3 model by A-FGSM, A-MI-FGSM and A-DIM with the kernel length ranging from 1 to 21 with a granularity 2. In Fig. 3, we show the success rates against five defense models—

Table 4: The success rates (%) of black-box attacks against eight defenses. The adversarial examples are crafted for the ensemble of Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 using FGSM, A-FGSM, MI-FGSM, A-MI-FGSM, DIM and A-DIM.

Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NIPS-r3
FGSM	27.5	23.7	13.4	4.9	13.8	38.1	30.0	19.8
A-FGSM	**39.1**	**38.8**	**31.6**	**29.9**	**31.2**	**43.3**	**39.8**	**33.9**
MI-FGSM	50.5	48.3	32.8	38.6	32.8	67.7	50.1	43.9
A-MI-FGSM	**76.4**	**74.4**	**69.6**	**73.3**	**68.3**	**77.2**	**72.1**	**71.4**
DIM	66.0	63.3	45.9	57.7	51.7	82.5	64.1	63.7
A-DIM	**84.8**	**82.7**	**78.0**	**82.6**	**81.4**	**83.4**	**79.8**	**83.1**



(a) A-FGSM (b) A-MI-FGSM (c) A-DIM

Figure 3: The success rates (%) of the adversarial examples generated for Inc-v3 agasinst IncRes-v2$_{ens}$, HGD, R&P, TVM and NIPS-r3, with the kernel length ranging from 1 to 21.

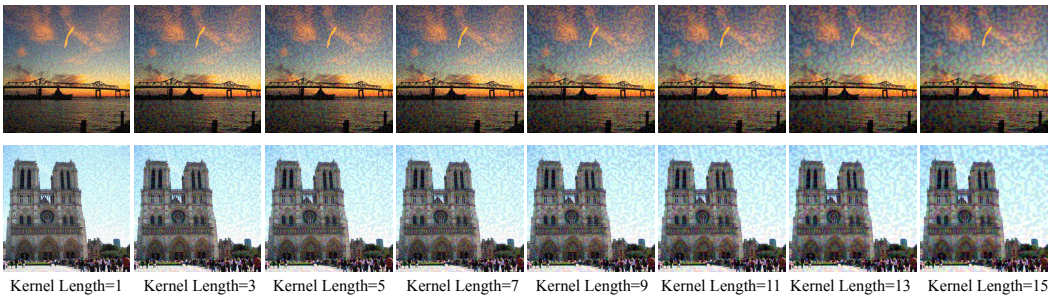


Figure 4: The adversarial examples generated for Inc-v3 by A-FGSM with different kernel sizes.

IncRes-v2$_{ens}$, HGD, R&P, TVM and NIPS-r3. The success rate continues increasing at first, and turns to remain stable after the kernel size exceeds 15×15.

We also show the adversarial images generated for the Inc-v3 model by A-FGSM with different kernel sizes in Fig. 4. Due to the smooth effect given by the kernel, we can see that the adversarial perturbations are smoother when using a bigger kernel.

5 Conclusion

In this paper, we propose an attention-invariant attack method to mitigate the attention shift phenomenon and generate more transferable adversarial examples against the defense models. Our method optimizes an adversarial image by using a set of shifted images. Based on an assumption, our method is simply implemented by convolving the gradient with a pre-defined kernel, and can be integrated into any gradient-based attack methods. We conduct experiments to validate the effectiveness of the proposed method. Our best attack A-DIM, the combination of the proposed attention-invariant method and diverse input iterative method (Xie et al., 2018b), can fool eight state-of-the-art defenses at an 82% success rate on average, where the adversarial examples are generated against four normally trained models. The results identify the vulnerability of the current defenses, which raises security issues for the development of more robust deep learning models.

REFERENCES

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

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018b.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Th European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 387–402, 2013.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Alex Graves, Abdel Rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. In *ICLR*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016b.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *Handbook of Brain Theory and Neural Networks*, 1995.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *ICLR*, 2017.

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*, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *ICLR*, 2018.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *ICLR*, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, 2018.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *ICLR*, 2018a.

Cihang Xie, Zhishuai Zhang, Jianyu Wang, Yuyin Zhou, Zhou Ren, and Alan Yuille. Improving transferability of adversarial examples with input diversity. 2018b.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

APPENDIX

We further show the results of the proposed attention-invariant attack method for white-box attacks and black-box attacks against normally trained models. We adopt the same settings for attacks. We also generate adversarial examples for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using FGSM, A-FGSM, MI-FGSM, A-MI-FGSM, DIM and A-DIM. For the attention-invariant based attacks, we set the kernel size as 7×7 since that the normally trained models have similar attentions. We then use these adversarial examples to attack six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16 and Res-v1-152. The results are shown in Table 5, Table 6 and Table 7. The attention-invariant based attacks get better results in most cases than the baseline attacks.

Table 5: The success rates (%) of adversarial attacks against six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16 and Res-v1-152. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using FGSM and A-FGSM. * indicates the white-box attacks.

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2-152	VGG-16	Res-v1-152
Inc-v3	FGSM	**79.6***	35.9	30.6	30.2	49.7	36.3
	A-FGSM	75.4*	**37.3**	**32.1**	**34.1**	**62.0**	**44.9**
Inc-v4	FGSM	43.1	**72.6***	32.5	34.3	50.7	37.7
	A-FGSM	**45.3**	68.1*	**33.7**	**35.4**	**63.3**	**46.2**
IncRes-v2	FGSM	44.3	36.1	**64.3***	31.9	49.4	38.6
	A-FGSM	**49.7**	**41.5**	63.7*	**40.1**	**64.2**	**46.7**
Res-v2-152	FGSM	40.1	34.0	30.3	**81.3***	50.5	40.8
	A-FGSM	**46.4**	**39.3**	**33.4**	78.9*	**64.7**	**50.4**

Table 6: The success rates (%) of adversarial attacks against six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16 and Res-v1-152. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using MI-FGSM and A-MI-FGSM. * indicates the white-box attacks.

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2-152	VGG-16	Res-v1-152
Inc-v3	MI-FGSM	97.8*	47.1	46.4	38.7	50.3	38.1
	A-MI-FGSM	**97.9***	**52.4**	**47.9**	**41.1**	**63.4**	**48.1**
Inc-v4	MI-FGSM	67.1	**98.8***	54.3	47.0	58.5	43.2
	A-MI-FGSM	**68.6**	**98.8***	**55.3**	**47.7**	**69.0**	**51.3**
IncRes-v2	MI-FGSM	74.8	64.8	**100.0***	54.5	59.3	50.8
	A-MI-FGSM	**76.1**	**69.5**	**100.0***	**59.6**	**74.4**	**61.5**
Res-v2-152	MI-FGSM	54.2	48.1	44.3	**97.5***	52.6	48.7
	A-MI-FGSM	**55.6**	**50.9**	**45.1**	97.4*	**65.6**	**59.6**

Table 7: The success rates (%) of adversarial attacks against six normally trained models—Inc-v3, Inc-v4, IncRes-v2, Res-v2-152, VGG-16 and Res-v1-152. The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 respectively using DIM and A-DIM. * indicates the white-box attacks.

	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-v2-152	VGG-16	Res-v1-152
Inc-v3	DIM	98.3*	73.8	67.8	58.4	62.5	49.3
	A-DIM	**98.5***	**75.2**	**69.2**	**59.0**	**74.3**	**59.1**
Inc-v4	DIM	**81.8**	98.2*	**74.2**	**65.1**	65.5	51.4
	A-DIM	80.7	**98.7***	73.2	62.7	**77.4**	**59.8**
IncRes-v2	DIM	86.1	83.5	**99.1***	73.5	67.9	62.7
	A-DIM	**86.4**	**85.5**	98.8*	**76.3**	**79.3**	**72.2**
Res-v2-152	DIM	**77.0**	**77.8**	**73.5**	**97.4***	67.4	67.8
	A-DIM	**77.0**	73.9	73.2	97.2*	**78.4**	**77.8**