NESTEROV ACCELERATED GRADIENT AND SCALE INVARIANCE FOR ADVERSARIAL ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning models are vulnerable to adversarial examples crafted by applying human-imperceptible perturbations on benign inputs. However, under the black-box setting, most existing adversaries often have a low transferability to attack other defense models. In this work, from the perspective of regarding the adversarial example generation as an optimization process, we propose two new methods to improve the transferability of adversarial examples, namely Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) and Scale-Invariant attack Method (SIM). NI-FGSM aims to adapt Nesterov accelerated gradient into the iterative attacks so as to effectively look ahead and avoid the "missing" of the global maximum. While SIM is based on our discovery on the scale-invariant property of deep learning models, for which we leverage to optimize the adversarial perturbations over the scale copies of the input images so as to avoid "overfitting" on the white-box model being attacked and generate more transferable adversarial examples. NI-FGSM and SIM can be naturally integrated to build a robust gradient-based attack to generate more transferable adversarial examples against the defense models. Empirical results on ImageNet dataset and NIPS 2017 adversarial competition demonstrate that our attack methods exhibit higher transferability and achieve higher attack success rates than state-of-the-art gradient-based attacks.

1 INTRODUCTION

Deep learning models have been shown to be vulnerable to adversarial examples (Goodfellow et al., 2014; Szegedy et al., 2014), which are generated by applying human-imperceptible perturbations on the benign input to result in misclassification. In addition, adversarial examples have an intriguing property of transferability in that adversarial examples crafted by the current model can also fool other unknown models. As adversarial examples can help to identify the robustness of models (Arnab et al., 2018), as well as improve the robustness of models by adversarial training (Goodfellow et al., 2014), learning how to generate robust adversarial examples with high transferability is important and has gained increasing attentions in the literature.

Several gradient-based attacks have been proposed to generate adversarial examples, such as one-step attacks (Goodfellow et al., 2014) and iterative attacks (Kurakin et al., 2016; Dong et al., 2018). Under the white-box setting, with the knowledge of the current model, existing attacks can achieve high success rates. However, they often exhibit low success rates under the black-box setting, especially for models with defense mechanism like adversarial training (Madry et al., 2017; Song et al., 2019) and input modification (Liao et al., 2018; Xie et al., 2018). Under the black-box setting, most existing attacks fail to generate adversarial examples with great transferability.

In this work, by regarding the adversarial example generation process as an optimization process, we propose two new methods to improve the transferability of adversarial examples: Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) and Scale-Invariant attack Method (SIM).

- Inspired by the fact that Nesterov accelerated gradient (Nesterov, 1983) is superior to momentum for conventionally optimization (Sutskever et al., 2013), we adapt Nesterov accelerated gradient into the iterative gradient-based attack so as to effectively look ahead and avoid the "missing" of the global maximum. We expect that NI-FGSM could replace the momentum iterative gradient-based method (Dong et al., 2018) in the gradient accumulating portion and yield higher performance.

- Besides, we discover that deep learning models have the *scale-invariant* property, and propose a Scale-Invariant attack Method (SIM) to improve the transferability of adversarial examples by optimizing the adversarial perturbations over the scale copies of the input images. SIM can avoid "overfitting" on the white-box model being attacked and generate more transferable adversarial examples to other black-box models.
- We found that combining our NI-FGSM and SIM with existing gradient-based attack methods (e.g., diverse input method (Xie et al., 2019)) can further boost the attack success rates of adversarial examples.

Extensive experiments on the ImageNet dataset (Russakovsky et al., 2015) show that our methods attack both normally trained models and adversarially trained models with higher attack success rates than existing baseline attacks. Our best attack method, SI-NI-TI-DIM (Scale-Invariant Nesterov Iterative FGSM integrated with translation-invariant diverse input method), reaches an average success rate of 93.5% against adversarial trained models under the black-box setting. For further demonstration, we evaluate our methods by attacking the top-3 defense solutions from NIPS 2017 adversarial competition (Kurakin et al., 2018). The results show that our attack method can generate adversarial examples with higher transferability than the top-1 attack solution of the NIPS2017 adversarial competition and the current strongest attacks.

2 Preliminary

2.1 Notation

Let x and y^{true} be a benign image and the corresponding true label, respectively. Let $J(x, y^{true})$ be the loss function of the classifier (e.g. the cross-entropy loss). Let x^{adv} be the adversarial example of the benign image x. The goal of the non-targeted adversaries is to search an adversarial example x^{adv} to maximize the loss $J(x^{adv}, y^{true})$ in the ℓ_p norm bounded perturbations. To align with previous works, we focus on $p = \infty$ in this work to measure the distortion between x^{adv} and x. That is $\left\|x^{adv} - x\right\|_\infty \le \epsilon$, where ϵ is the magnitude of adversarial perturbations.

2.2 Attack Methods

Here we introduce several advanced gradient-based methods for generating adversarial examples.

Fast Gradient Sign Method (FGSM). FGSM (Goodfellow et al., 2014) generates an adversarial example x^{adv} by maximizing the loss function $J(x^{adv}, y^{true})$ with one-step update as:

$$x^{adv} = x + \epsilon \cdot \mathrm{sign}(\nabla_x J(x, y^{true})), \tag{1}$$

where $\mathrm{sign}(\cdot)$ function restricts the perturbation in the L_∞ norm bound.

Iterative Fast Gradient Sign Method (I-FGSM). Kurakin et al. (2016) extend FGSM to an iterative version by applying FGSM in iterations with a small step size α:

$$\begin{aligned} x_0 &= x, \\ x_{t+1}^{adv} &= \mathrm{Clip}_x^\epsilon\{x_t^{adv} + \alpha \cdot \mathrm{sign}(\nabla_x J(x_t^{adv}, y^{true}))\}, \end{aligned} \tag{2}$$

where $\mathrm{Clip}_x^\epsilon(\cdot)$ function restricts the generated adversarial examples x_t^{adv} to be within the ϵ-ball of x.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM). Dong et al. (2018) integrate momentum into the iterative attack and lead to higher transferability of adversarial examples. Their update procedure is formalized as follows:

$$\begin{aligned} g_{t+1} &= \mu \cdot g_t + \frac{\nabla_x J(x_t^{adv}, y^{true})}{\left\|\nabla_x J(x_t^{adv}, y^{true})\right\|_1}, \\ x_{t+1}^{adv} &= \mathrm{Clip}_x^\epsilon\{x_t^{adv} + \alpha \cdot \mathrm{sign}(g_{t+1})\}, \end{aligned} \tag{3}$$

where g_t is the accumulated gradient at iteration t, and μ is the decay factor of g_t.

Diverse Input Method (DIM). Xie et al. (2019) optimize the adversarial perturbations over the diverse transformation of the input image at each iteration. The transformations include the random resizing and the random padding. DIM can be naturally integrated into other gradient-based attacks to further improve the transferability of adversarial examples.

Translation-Invariant Method (TIM). Instead of optimizing the adversarial perturbations on a single image, Dong et al. (2019) use a set of translated images to optimize the adversarial perturbations. They further develop an efficient algorithm to calculate the gradients by convolving the gradient at untranslated images with a kernel matrix. TIM can also be naturally integrated with other gradient-based attack methods. TI-DIM, the combination of TIM and DIM, is the current strongest black-box attack methods.

2.3 DEFENSE METHODS

Various defense methods have been proposed to against adversarial examples, which can fall into the following two categories.

Adversarial training. One popular defense method is *adversarial training* (Goodfellow et al., 2014; Szegedy et al., 2014), which augments the training data by the adversarial examples during the training process. Tramr et al. (2018) propose *ensemble adversarial training* by augmenting the training data with perturbations transferred from various models to further improve the robustness against the black-box attacks.

Input modification. The second category of defense methods aims to mitigate the effects of adversarial perturbations by modifying the input data. Guo et al. (2018) discover that there exists a range of image transformations that have the potential to remove adversarial perturbations while preserving the visual information of the images. Xie et al. (2018) mitigate the adversarial effects through random transformations. Liao et al. (2018) propose high-level representation guided denoiser to purify the adversarial examples.

3 METHODOLOGY

3.1 MOTIVATION

Similar with the process of training deep neural networks, the process of generating adversarial examples can also be viewed as an optimization problem. In the optimizing phase, the white-box model being attacked to optimize the adversarial examples can be viewed as the training data on the training process. And the adversarial examples can be viewed as the training parameters of the model. Then in the testing phase, the black-box models to evaluate the adversarial examples can be viewed as the testing data of the model.

From the perspective of the optimization, the transferability of the adversarial examples is similar with the generalization ability of the training model. Thus, we can migrate the methods used to improve the generalization of models to the generating of adversarial examples, so as to improving the transferability of adversarial examples.

Many methods have been proposed to improve the generalization ability of the deep learning models, which can be split to two aspects: (1) better optimization algorithm, such as Adam optimizer(Kingma & Ba, 2014); (2) data augmentation (Simonyan & Zisserman, 2014). Correspondingly, methods to improve the transferability of adversarial examples can also be split to two aspects: (1) better optimization algorithm, such as MI-FGSM that applies the idea of momentum; (2) model augmentation (i.e., ensemble attack on multiple models), such as the work of Dong et al. (2018) that considers to attack multiple models simultaneously. Based on above analysis, we aim to improve the transferability of adversarial examples by applying the idea of Nesterov accelerated gradient for *optimization* and using a set of scaled images to achieve *model augmentation*.

3.2 NESTEROV ITERATIVE FAST GRADIENT SIGN METHOD

Nesterov Accelerated Gradient (NAG) (Nesterov, 1983) is a slight variation of normal gradient descent that can speed up the training and improve convergence significantly. NAG can be viewed as an improved momentum method, which can be expressed as:

$$\begin{aligned} v_{t+1} &= \mu \cdot v_t + \nabla_{\theta_t} J(\theta_t - \alpha \cdot \mu \cdot v_t), \\ \theta_{t+1} &= \theta_t - \alpha \cdot v_{t+1}. \end{aligned} \tag{4}$$

Compared to momentum (Polyak, 1964), the anticipatory update of NAG can prevent from going too fast and lead to the increased responsiveness, so as to avoid the “missing” of the global maximum. Therefore, NAG is expected to have better performance than momentum. Sutskever et al. (2013) provide empirical evidence that NAG is superior to momentum for conventionally difficult optimization problems.

We naturally integrate NAG into the iterative gradient-based attack to build a robust adversarial attack, which we refer to as NI-FGSM (Nesterov Iterative Fast Gradient Sign Method). Specifically, we make a jump in the direction of previous accumulated gradients before computing the gradients in each iteration. Start with $g_0 = 0$, the update procedure of NI-FGSM can be formalized as follows:

$$x_t^{nes} = x_t^{adv} + \alpha \cdot \mu \cdot g_t, \tag{5}$$

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t^{nes}, y^{true})}{\|\nabla_x J(x_t^{nes}, y^{true})\|_1}, \tag{6}$$

$$x_{t+1}^{adv} = \mathrm{Clip}_x^{\epsilon}\{x_t^{adv} + \alpha \cdot \mathrm{sign}(g_{t+1})\}, \tag{7}$$

where g_t denotes the accumulated gradients at iteration t, and μ denotes the decay factor of g_t.

3.3 Scale-Invariant Attack Method

Other than considering a better optimization algorithm for the adversaries, we can also improve the transferability of adversarial examples by *model augmentation*. We first introduce a formal definition of loss-preserving transformation and model augmentation as follows.

Definition 1 ***Loss-preserving Transformation.*** *Given an input x with its ground-truth label y^{true} and a classifier $f(x) : x \in \mathcal{X} \to y \in \mathcal{Y}$ with the cross-entropy loss $J(x, y)$, if there exists an input transformation $\mathcal{T}(\cdot)$ that satisfies $J(\mathcal{T}(x), y^{true}) \approx J(x, y^{true})$ for any $x \in \mathcal{X}$, we say $\mathcal{T}(\cdot)$ is a loss-preserving transformation.*

Definition 2 ***Model Augmentation.*** *Given an input x with its ground-truth label y^{true} and a model $f(x) : x \in \mathcal{X} \to y \in \mathcal{Y}$ with the cross-entropy loss $J(x, y)$, if there exists a loss-preserving transformation $\mathcal{T}(\cdot)$, then we derive a new model by $f'(x) = f(\mathcal{T}(x))$ from the original model f. we define such derivation of models as model augmentation.*

Intuitively, similar to the generalization of models that can be improved by feeding more training data, the transferability of adversarial examples can be improved by attack more models simultaneously. Dong et al. (2018) enhance the gradient-based attack by attacking an ensemble of models. However, their approach requires training a set of different models to attack, which has a large computational cost. Instead, in this work, we derive an ensemble of models from the original model by *model augmentation*, which is a simple way of obtaining multiple models via the loss-preserving transformation.

To get the loss-preserving transformation, we discover that deep neural networks may have the scale-invariant property, besides the translation invariance. Specifically, the loss values are similar for the original and the scaled images on the same model, which is empirically validated in Section 4.2. Thus, the scale transformation can be served as a model augmentation method. Driven by the above analysis, we propose a Scale-Invariant attack Method (SIM), which optimizes the adversarial perturbations over the scale copies of the input image:

$$\begin{aligned} &\arg\max_{x^{adv}} \frac{1}{m} \sum_{i=0}^{m} J(S_i(x^{adv}), y^{true}), \\ &\text{s.t.} \left\| x^{adv} - x \right\|_\infty \leq \epsilon, \end{aligned} \tag{8}$$

where $S_i(x) = x/2^i$, which denotes the scale copy of the input image x with the scale factor $1/2^i$ and m denotes the number of the scale copies. With SIM, instead of training a set of models to attack, we can effectively achieve ensemble attacks on multiple models by model augmentation. More importantly, it can help avoid "overfitting" on the white-box model being attacked and generate more transferable adversarial examples.

3.4 Attack Algorithm

For the gradient processing of crafting adversarial examples, NI-FGSM introduce a better optimization algorithm to stabilize and correct the update directions at each iteration. For the ensemble attack of crafting adversarial examples, SIM introduces *model augmentation* to derive multiple models to attack from a single model. Thus, NI-FGSM and SIM can be naturally combined to build a stronger attack, which we refer to SI-NI-FGSM (Scale-Invariant Nesterov Iterative Fast Gradient Sign Method). The algorithm of SI-NI-FGSM attack is summarized in Algorithm 1.

Algorithm 1 SI-NI-FGSM

Input: A clean example x with ground-truth label y^{true}; a classifier f with loss function J;
Input: Perturbation size ϵ; maximum iterations T; number of scale copies m and decay factor μ.
Output: An adversarial example x^{adv}

$\alpha = \epsilon/T$
$g_0 = 0; x_0^{adv} = x$
for $t = 0$ to $T-1$ **do**
    $g = 0$
    Get x_t^{nes} following Eq.(5) ▷ make a jump in the direction of previous accumulated gradients
    for $i = 0$ to $m-1$ **do** ▷ sum the gradients over the scale copies of the input image
        Get the gradients by $\nabla_x J(S_i(x_t^{nes}), y^{true})$
        Sum the gradients as $g = g + \nabla_x J(S_i(x_t^{nes}), y^{true})$
    Get average gradients as $g = \frac{1}{m} \cdot g$
    Update g_{t+1} by $g_{t+1} = \mu \cdot g_t + \frac{g}{\|g\|_1}$
    Update x_{t+1}^{adv} by Eq.(7)
return $x^{adv} = x_T^{adv}$

In addition, SI-NI-FGSM can be integrated with DIM (Diverse Input Method), TIM (Translation-Invariant Method) and TI-DIM (Translation-Invariant with Diverse Input Method) as SI-NI-DIM, SI-NI-TIM and SI-NI-TI-DIM, respectively, to further boost the transferability of adversarial examples. The detail algorithms for these attack methods are provided in Appendix A.

4 Experimental Results

In this section, we provide experimental evidence on the advantage of the proposed methods. We first provide the setting of the experiments in Section 4.1. Then we explore the scale-invariant property of deep learning model in Section 4.2. Then we perform ablation studies on our method and compare it with MI-FGSM. Finally, we show that our method can be integrated with the current attacks to boost the average attack success rate under single-model and multi-model settings. Codes are available online[1].

4.1 Experimental Setup

Dataset. We randomly choose 1000 images belonging to the 1000 categories from the ILSVRC 2012 validation set, which are almost correctly classified by all the testing models.

Models. We consider seven different models. Four of them are 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-v2-101 (Res-101) (He et al., 2016). The others are adversarially trained models, Inc-v3$_{\text{ens3}}$, Inc-v3$_{\text{ens4}}$, and IncRes-v2$_{\text{ens}}$ (Tramr et al., 2018).

[1] https://drive.google.com/drive/folders/1NNgw8GJC624aKfEzOp1oTlOupETEpFr4?usp=sharing

Baselines. We first validate the impact of each component of SI-NI-FGSM by comparing with MI-FGSM (Dong et al., 2018). Then we integrate SI-NI-FGSM with DIM (Xie et al., 2019), TIM, and TI-DIM (Dong et al., 2019), to further show the effectiveness of SI-NI-FGSM. We denote the attacks combined with our SI-NI-FGSM as SI-NI-DIM, SI-NI-TIM, and SI-NI-TIM-DIM, respectively.

Hyper-parameters. For the hyper-parameters, we follow the settings in (Dong et al., 2018) with the maximum perturbation as $\epsilon = 16$, the number of iteration as $T = 16$, and the step size as $\alpha = 1.6$. For MI-FGSM, we adopt the default decay factor $\mu = 1.0$. For DIM, the transformation probability is set to 0.5. For TIM, we adopt the Gaussian kernel and the size of the kernel is set to 7×7. For SI-NI-FGSM, we set the number of the scale copies m as 5.

4.2 Scale-Invariant Property

To validate the scale-invariant property of deep neural networks, we randomly choose 1,000 original images from the ImageNet dataset and keep the scale size in the range of $[0.1, 2.0]$ with a step size 0.1. Then we feed the scaled images into the models, including Inc-v3, Inc-v4, IncRes-2, and Res-101, to get the average loss value over 1,000 images.

As shown in Figure 1, we can easily observe that the loss curves are smooth and stable when the scale size is in the range of $[0.1, 1.3]$. That is, the loss values would be very similar for the original and scaled images. So we assume that the scale-invariant property is held within the range $[0.1, 1.3]$. In our attack, we leverage the scale-invariant property to optimize the adversarial perturbations over the scale copies of the input images.

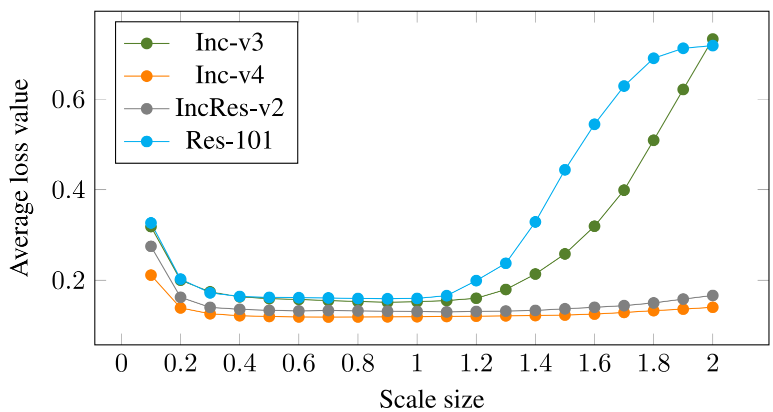

Figure 1: **The curves of average loss value for Inc-v3, Inc-v4, IncRes-v2 and Res-101 given the scaled images at each scale size.** The results are averaged over 1000 images.

4.3 NI-FGSM Vs MI-FGSM

We compare our NI-FGSM with the baseline attack MI-FGSM (Dong et al., 2018) and dissect the impact of SIM by integrating SIM into NI-FGSM and MI-FGSM. As shown in Table 1, as compared with MI-FGSM, NI-FGSM achieves higher attack success rate under the black-box setting. The attack success rates of SI-NI-FGSM and SI-MI-FGSM against the defenses are improved by a large margin, which we can directly attribute to SIM. Some adversarial examples generated by our proposed SI-NI-FGSM are shown in Appendix B.

4.4 Attack Single Model

In this subsection, we extend our SI-NI-FGSM by combining with the baselines (TIM, DIM and TI-DIM), and compare the black-box attack success rates of our extensions with the baselines under the single model setting. As shown in Table 2, we observe that our extension methods consistently outperform the baseline attacks by $10\% \sim 35\%$ under the black-box setting, and achieve nearly 100% success rates under the white-box setting. It indicates that SI-NI-FGSM can serve as a powerful approach to improve the transferability of adversarial examples.

Table 1: **The attack success rate (%) of adversarial attacks against the models.** The adversarial examples are crafted for Inc-v3 using MI-FGSM, NI-FGSM, SI-MI-FGSM, and SI-NI-FGSM, and $*$ indicates the white-box model being attacked.

Attack	Inc-v3*	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{\text{ens3}}$	Inc-v3$_{\text{ens4}}$	IncRes-v2$_{\text{ens}}$	**Average**
MI-FGSM	**100.0**	42.7	41.8	35.2	14.7	12.5	6.1	35.1
NI-FGSM(**Ours**)	**100.0**	52.6	51.4	41.0	12.9	12.8	6.4	39.6
SI-MI-FGSM(**Ours**)	**100.0**	69.1	67.0	63.1	**32.0**	**31.0**	17.2	54.2
SI-NI-FGSM(**Ours**)	**100.0**	**76.0**	**73.3**	**67.6**	31.6	30.0	**17.4**	**56.6**

Table 2: **The attack success rates (%) of adversarial attacks against seven models under single-model setting.** The adversarial examples are crafted for Inc-v3, Inc-v4, IncRes-v2, and Res-101 respectively. * indicates the white-box attacks.

(a) **Comparison of TIM and the extension SI-NI-TIM.**

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{\text{ens3}}$	Inc-v3$_{\text{ens4}}$	IncRes-v2$_{\text{ens}}$
Inc-v3	TIM	**100.0***	47.8	42.8	39.5	24.0	21.4	12.9
	SI-NI-TIM (**Ours**)	**100.0***	**77.2**	**75.8**	**66.5**	**51.8**	**45.9**	**33.5**
Inc-v4	TIM	58.5	99.6*	47.5	43.2	25.7	23.3	17.3
	SI-NI-TIM (**Ours**)	**83.5**	**100.0***	**76.6**	**68.9**	**57.8**	**54.3**	**42.9**
IncRes-v2	TIM	62.0	56.2	97.5*	51.3	32.8	27.9	21.9
	SI-NI-TIM (**Ours**)	**86.4**	**83.2**	**99.5***	**77.2**	**66.1**	**60.2**	**57.1**
Res-101	TIM	59.0	53.6	51.8	99.3*	36.8	32.2	23.5
	SI-NI-TIM (**Ours**)	**78.3**	**74.1**	**73.0**	**99.8***	**58.9**	**53.9**	**43.1**

(b) **Comparison of DIM and the extension SI-NI-DIM.**

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{\text{ens3}}$	Inc-v3$_{\text{ens4}}$	IncRes-v2$_{\text{ens}}$
Inc-v3	DIM	98.7*	67.7	62.9	54.0	20.5	18.4	9.7
	SI-NI-DIM (**Ours**)	**99.6***	**84.7**	**81.7**	**75.4**	**36.9**	**34.6**	**20.2**
Inc-v4	DIM	70.7	98.0*	63.2	55.9	21.9	22.3	11.9
	SI-NI-DIM (**Ours**)	**89.7**	**99.3***	**84.5**	**78.5**	**47.6**	**45.0**	**28.9**
IncRes-v2	DIM	69.1	63.9	93.6*	57.4	29.4	24.0	17.3
	SI-NI-DIM (**Ours**)	**89.7**	**86.4**	**99.1***	**81.2**	**55.0**	**48.2**	**38.1**
Res-101	DIM	75.9	70.0	71.0	98.3*	36.0	32.4	19.3
	SI-NI-DIM (**Ours**)	**88.7**	**84.2**	**84.4**	**99.3***	**52.4**	**48.0**	**33.2**

(c) **Comparison of TI-DIM and the extension SI-NI-TI-DIM.**

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{\text{ens3}}$	Inc-v3$_{\text{ens4}}$	IncRes-v2$_{\text{ens}}$
Inc-v3	TI-DIM	98.5*	66.1	63.0	56.1	38.6	34.9	22.5
	SI-NI-TI-DIM (**Ours**)	**99.6***	**85.5**	**80.9**	**75.7**	**61.5**	**56.9**	**40.7**
Inc-v4	TI-DIM	72.5	97.8*	63.4	54.5	38.1	35.2	25.3
	SI-NI-TI-DIM (**Ours**)	**88.1**	**99.3***	**83.7**	**77.0**	**65.0**	**63.1**	**49.4**
IncRes-v2	TI-DIM	73.2	67.5	92.4*	61.3	46.4	40.2	35.8
	SI-NI-TI-DIM (**Ours**)	**89.6**	**87.0**	**99.1***	**83.9**	**74.0**	**67.9**	**63.7**
Res-101	TI-DIM	74.9	69.8	70.5	98.7*	52.6	49.1	37.8
	SI-NI-TI-DIM (**Ours**)	**86.4**	**82.6**	**84.6**	**99.0***	**72.6**	**66.8**	**56.4**

4.5 Attack Ensemble of Models

Following the work of (Liu et al., 2016), we would like to show the performance of our methods when attacking multiple models simultaneously. Specifically, we attack an ensemble of normally trained models (including Inc-v3, Inc-v4, IncRes-v2 and Res-101) with equal ensemble weights using TIM, SI-NI-TIM, DIM, SI-NI-DIM, TI-DIM and SI-NI-TI-DIM, respectively.

As shown in Table 3, our methods improve the attack success rates across all experiments over the baselines. In general, our methods consistently outperform the baseline attacks by 10% $\sim$ 30% under the black-box setting. In particular, SI-NI-TI-DIM, the extension by combining SI-NI-FGSM and TI-DIM, can fool the adversarial trained models with an average success rate of 93.5%, which indicates current adversarial trained models provide few benefits for the robustness when meeting SI-NI-TI-DIM under the black-box setting.

Table 3: **The attack success rates (%) of adversarial attacks against seven models under multi-model setting.** * indicates the white-box models being attacked.

Attack	Inc-v3*	Inc-v4*	IncRes-v2*	Res-101*	Inc-v3$_{\mathrm{ens3}}$	Inc-v3$_{\mathrm{ens4}}$	IncRes-v2$_{\mathrm{ens}}$
TIM	99.9	99.3	99.3	99.8	71.6	67.0	53.2
SI-NI-TIM (**Ours**)	**100.0**	**100.0**	**100.0**	**100.0**	**93.2**	**90.1**	**84.5**
DIM	99.7	99.2	98.9	98.9	66.4	60.9	41.6
SI-NI-DIM (**Ours**)	**100.0**	**100.0**	**100.0**	**99.9**	**88.2**	**85.1**	**69.7**
TI-DIM	99.6	98.8	98.8	98.9	85.2	80.2	73.3
SI-NI-TI-DIM (**Ours**)	**99.9**	**99.9**	**99.9**	**99.9**	**96.0**	**94.3**	**90.3**

4.6 Comparison on NIPS 2017 Adversarial Competition

We consider to quantify the effectiveness of our methods by evaluating on the top-3 defense solutions from the NIPS 2017 adversarial competition, which are high-level representation guided denoiser (HGD, rank-1) (Liao et al., 2018), input transformation though random resizing and padding (R&P, rank-2) (Xie et al., 2018), and rank-3 submission (NIPS-r3)[2].

We compare our SI-NI-TI-DIM with MI-FGSM (Dong et al., 2018), which is the top-1 attack solution in the NIPS competition, and TI-DIM (Dong et al., 2019), which is state-of-the-art attack method. We first generate adversarial examples for the ensemble models, including Inc-v3, Inc-v4, IncRes-v2, and Res-101 by using MI-FGSM, TI-DIM, and SI-NI-TI-DIM, respectively. Then, we evaluate the adversarial examples by attacking the top-3 defense solutions.

As shown in Table 4, our method SI-NI-TI-DIM achieves an average attack success rate of 93.9%, which outperforms the state-of-the-art attack method by a large margin of 13.6%. The results show that, only depending on the transferability of adversarial examples, SI-NI-TI-DIM can fool not only the adversarial trained models but also other strong defense mechanism, raising new security issues for the development of more robust deep learning models.

Table 4: **The attack success rates (%) of adversarial attacks against top-3 defense solutions from NIPS 2017 adversarial competition.**

Attack	HGD	R&P	NIPS-r3	**Average**
MI-FGSM	36.9	29.3	40.8	35.7
TI-DIM	84.8	75.3	80.7	80.3
SI-NI-TI-DIM (**Ours**)	**96.1**	**91.3**	**94.4**	**93.9**

5 Conclusion

In this paper, by an analogy with the training of deep models, we regard the process of generating adversarial examples as an optimization problem, where the current white-box model is the training data, other black-box models are the testing data, and the generated examples are the parameters to be trained. We wish the generated examples have a good transferability and design two methods for a better optimization algorithm and data argumentation, respectively.

Specifically, we propose two new attack methods, NI-FGSM and SIM, to improve the transferability of adversarial examples. NI-FGSM adopts the Nesterov accelerated gradient method into the gradient-based attack, and SIM aims to achieve *model augmentation* by leverage the scale-invariant property of models. NI-FGSM and SIM can be naturally combined to build a more robust attack called SI-NI-FGSM. Moreover, by integrating SI-NI-FGSM with baseline attacks, we can further improve the transferability of adversarial examples. Extensive experiments demonstrate that our methods not only yield higher success rates on the adversarial trained models but also break other strong defense mechanism with nearly 94% attack success rate. We believe our methods could serve as a strong benchmark for evaluating the robustness of various defense models.

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

REFERENCES

Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 888–897, 2018.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

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

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, pp. 195–231. Springer, 2018.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

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.

Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, volume 269, pp. 543–547, 1983.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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

Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Improving the generalization of adversarial training with domain adaptation. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147, 2013.

Christian Szegedy, Google Inc, Wojciech Zaremba, Ilya Sutskever, Google Inc, Joan Bruna, Dumitru Erhan, Google Inc, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations, Workshop Track*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

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

Florian Tramr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

A Details of the Algorithms

The algorithm of SI-NI-TI-DIM attack is summarized in Algorithm 2. We can get the SI-NI-DIM attack algorithm by removing Step 10 of Algorithm 2, and get the SI-NI-TIM attack algorithm by removing $T(\cdot; p)$ in Step 7 of Algorithm 2.

Algorithm 2 SI-NI-TI-DIM

Input: A clean example x with ground-truth label y^{true}; a classifier f with loss function J;
Input: Perturbation size ϵ; maximum iterations T; number of scale copies m and decay factor μ.
Output: An adversarial example x^{adv}

1: $\alpha = \epsilon / T$
2: $g_0 = 0; x_0^{adv} = x$
3: **for** $t = 0$ to $T - 1$ **do**
4: $\quad g = 0$
5: $\quad$ Get x_t^{nes} by Eq.(5) $\triangleright$ make a jump in the direction of previous accumulated gradients
6: $\quad$ **for** $i = 0$ to $m - 1$ **do** $\triangleright$ sum the gradients over the scale copies of the input image
7: $\quad\quad$ Get the gradients by $\nabla_x J(T(S_i(x_t^{nes}); p), y^{true})$ $\triangleright$ apply random resizing and padding to the inputs with the probability p
8: $\quad\quad$ Sum the gradients as $g = g + \nabla_x J(T(S_i(x_t^{nes}); p), y^{true})$
9: $\quad$ Get average gradients as $g = \frac{1}{m} \cdot g$
10: $\quad$ Convolve the gradients by $g = \boldsymbol{W} * g$ $\triangleright$ convolve gradient with the pre-defined kernel $\boldsymbol{W}$
11: $\quad$ Update g_{t+1} by $g_{t+1} = \mu \cdot g_t + \frac{g}{\|g\|_1}$
12: $\quad$ Update x_{t+1}^{adv} by Eq.(7)
13: **return** $x^{adv} = x_T^{adv}$

B Visualization of Adversarial Examples

We visualize 12 randomly selected benign images and their corresponding adversarial images in Figure 2. The adversarial images are crafted for Inc-v3 using the proposed SI-NI-FGSM. We see that these generated adversarial perturbations are human imperceptible.

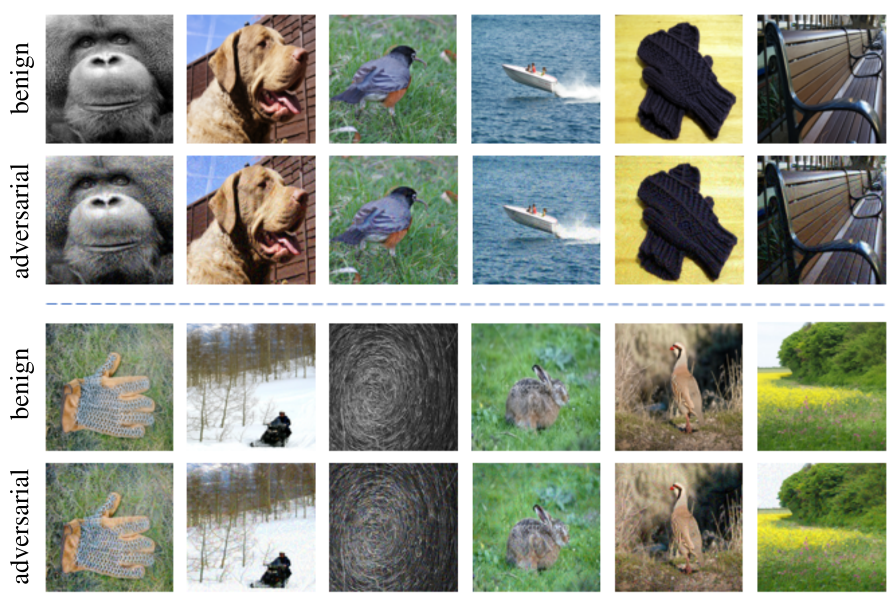

Figure 2: Visualization of randomly picked benign images and their corresponding adversarial images, crafted on Inc-v3 using SI-NI-FGSM.