Boosting Adversarial Transferability through Enhanced Momentum

Xiaosen Wang*[1], Jiadong Lin*[1]
{xiaosen,jdlin}@hust.edu.cn
Han Hu[2], Jingdong Wang[2]
{hanhu,jingdw}@microsoft.com
Kun He†[1]
brooklet60@hust.edu.cn

[1] School of Computer Science and Technology, Huazhong University of Science and Technology
[2] Microsoft Research Asia
*Equal contribution
†Corresponding author

Abstract

Deep learning models are known to be vulnerable to adversarial examples crafted by adding human-imperceptible perturbations on benign images. Many existing adversarial attacks have achieved great white-box attack performance, but exhibit low transferability when attacking other models. Various momentum iterative gradient-based methods are shown to be effective to improve the adversarial transferability. In what follows, we propose an enhanced momentum iterative gradient-based method to further enhance the adversarial transferability. Specifically, instead of only accumulating the gradient during the iterative process, we additionally accumulate the average gradient of the data points sampled in the gradient direction of the previous iteration so as to stabilize the update direction and escape from poor local maxima. Extensive experiments on the standard ImageNet dataset demonstrate that our method could improve the adversarial transferability of momentum-based methods by a large margin of 11.1% on average. Moreover, by incorporating with various input transformations, the adversarial transferability could be further improved significantly. We also attack several extra advanced defense models in the ensemble-model setting, and the enhancements are at least 7.8% on average.

1 Introduction

With the impressive performance of deep neural networks (DNNs) [4, 10, 12, 13, 21], the vulnerability to adversarial examples [11, 31], which are indistinguishable from legitimate ones by adding tiny perturbations but lead to erroneous predictions, has raised serious concerns in security-sensitive applications, *e.g.* self-driving automobile [2], face verification [30] *etc*. This issue of DNNs has triggered two research directions, with one trying to improve the attack ability of adversarial examples [1, 2, 11, 14, 16, 35] and the other line studying to improve the robustness of neural networks against the adversaries [3, 17, 22, 37, 38, 41]. The two directions, namely *adversarial attack* and *adversarial defense*, usually act like spear and shield that the progress on one side can inspire the improvements of the other side.

For adversarial attack, numerous methods have been proposed in recent years, such as the one-step gradient-based attacks [11, 34], iterative gradient-based attacks [14, 22], and optimization-based attacks [2, 31]. Existing adversarial attacks often fall into the category

of white-box setting, where the adversary is capable to access all information about the target model. For the counterpart category of black-box attacks, adversarial transferability, *i.e.* the ability of adversarial examples generated on one model to mislead other models, is an important metric. Such property makes it possible to attack deep neural models without knowing any inner working mechanism in practice. Though white-box attacks achieve good attack performance, they often exhibit low transferability.

Recently, various methods are proposed to improve the transferability of white-box attacks, *e.g.* incorporating momentum into iterative gradient-based attacks [6, 18], ensemble-model attack [19], input transformations [6, 18, 36, 39] *etc*. Note that both ensemble-model attack and input transformations are based on existing gradient-based attacks. However, NI-FGSM, which exhibits the best transferability among existing momentum based attacks [18], can only achieve the average attack success rate of less than 52% in black-box setting, as shown in Table 1, indicating that the improvement of ensemble-model attack and transformation-based attack is rather limited.

In this work, inspired by momentum based attacks, we propose an enhanced momentum iterative fast gradient sign method (EMI-FGSM), to further promote the transferability. As shown in Figure 1, different from existing momentum based methods (*e.g.* MI-FGSM) that only accumulate the gradients of data points along the optimization path, EMI-FGSM additionally accumulates the gradients of data points sampled in the gradient direction of previous iteration. Such accumulation might help find more stable gradient direction, leading to better local maxima. Empirical evaluations show that EMI-FGSM achieves higher attack success rates in white-box setting and significantly higher transferability in black-box setting.

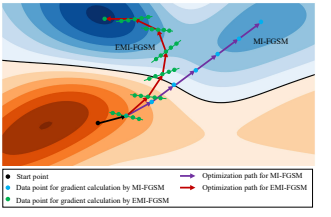


Figure 1: Illustration of optimization path of MI-FGSM [6] and EMI-FGSM. At each iteration, MI-FGSM accumulates the gradient of data point along the path, while EMI-FGSM accumulates the gradients of data points sampled in the gradient direction of previous iteration, which helps EMI-FGSM find better local maxima for higher transferability.

Moreover, EMI-FGSM is complementary to ensemble-model attack and input transformations. When integrated with these advanced methods, the enhanced momentum equipped methods can achieve significantly higher transferability on standard ImageNet dataset than SOTA baselines. When attacking seven advanced defenses that exhibit good effectiveness against transferability on ImageNet, our method combined with input transformations in ensemble-model setting achieves an average attack success rate of 86.6%, improving the transferability of existing advanced attacks by a clear margin of 7.8%.

2 Related Work

Given a classifier f and an input image x, where $f(x)$ outputs the prediction label of x. Let $J_f(x,y)$ denote the loss function of classifier f and $\mathcal{B}_\varepsilon(x) = \{x' : \|x - x'\|_p \leq \varepsilon\}$ denote the L_p-norm ball centered at x with radius ε and we focus on L_∞-norm as in previous works.

Adversarial attack can be formulated as $x^{adv} \in \mathcal{B}_\varepsilon(x)$ s.t. $f(x) \neq f(x^{adv})$. Based on the threat model, existing adversarial attacks can be roughly categorized into two settings: a) *white-box attack* allows full access to the threat model, *e.g.* model outputs, gradients and architectures, *etc*. b) *black-box attack* only allows access to the model outputs. Recent works also find that adversaries have good transferability [19, 26] across different models, *i.e.* the adversaries generated on one model can still fool other models, falling into black-box attack.

Existing white-box adversarial attacks [2, 14, 22, 23, 25] usually optimize the perturbation using the gradient and exhibit good attack performance but low transferability. To boost the transferability, several gradient-based adversarial attacks have been proposed. Dong *et al*. [5] propose to integrate momentum into iterative gradient-based attack. Lin *et al*. [18] adopt Nestorve's accelerated gradient for higher transferability. Liu *et al*. [19] show that ensemble-model attack which attacks multiple models simultaneously, can improve the transferability.

Recent works also find that input transformations can further enhance the tranferability. Diverse Input Method (DIM) [39] creates diverse input patterns by applying random resizing and padding to the input before feeding the image into the model for gradient calculation. Translation-Invariant Method (TIM) [6] optimizes the perturbation over an ensemble of the translated images by convolving the gradient at the untranslated image with a pre-defined kernel. Scale-Invariant Method (SIM) [18] optimizes the adversarial perturbation over m scale copies of the input to achieve higher transferability. Gao *et al*. [8] propose a patch-wise iterative method (PIM), which projects the excess noise into the surrounding field and could be integrated with existing iterative gradient-based attacks for more transferable adversaries. Some works [9, 15, 42] focus on crafting more transferable target adversarial examples.

Ensemble-model attack and input transformation can be combined with gradient-based methods to further improve the transferability. Our method is a new variation of gradient-based attack with higher transferability and can be integrated with ensemble-model attack and input transformation to achieve higher transferability.

3 Methodology

In this section, we first give an overview of gradient-based adversarial attacks, to which our method belongs. Then we provide detailed descriptions of the proposed Pre-gradient guided momentum Iterative FGSM (PI-FGSM) and Enhanced Momentum I-FGSM (EMI-FGSM).

3.1 Gradient-based Adversarial Attacks

Gradient-based adversarial attacks are typical methods for adversarial attacks.

Fast Gradient Sign Method (FGSM) [11] generates adversaries by a one-step update:

$$x^{adv} = x + \varepsilon \cdot \text{sign}(\nabla_x J_f(x,y)),$$

where $\text{sign}(\cdot)$ is the sign function and $\nabla_x J_f$ denotes the gradient of the loss function w.r.t. x.

Iterative Fast Gradient Sign Method (I-FGSM) [14] extends FGSM by iteratively applying the gradient update:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} J_f(x_t^{adv},y)),$$

where $x_1^{adv} = x$, $\alpha = \varepsilon/T$ is a small step size, and T is the number of iterations.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [5] proposes to integrate the momentum [27] into the iterative attack to achieve higher transferability:

$$g_t = \mu \cdot g_{t-1} + \frac{\nabla_{x_t^{adv}} J_f(x_t^{adv},y)}{\|\nabla_{x_t^{adv}} J_f(x_t^{adv},y)\|_1}, \quad x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_t),$$

where g_{t-1} is the accumulated gradient at $(t-1)$-th iteration with a decay factor μ and $g_0 = 0$.

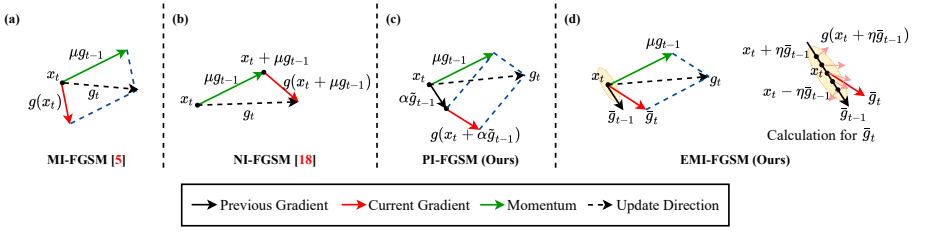


Figure 2: Illustration of gradient update at t-th iteration for various momentum based attack, where $g(x)$ denotes the gradient of x. (a) MI-FGSM [5] accumulates the gradient of x_t for update. (b) NI-FGSM [18] accumulates the gradient of $x_t + \mu g_{t-1}$ for update. (c) PI-FGSM accumulates the gradient of $x_t + \alpha\tilde{g}_{t-1}$ for update, where $\tilde{g}_{t-1}$ is the gradient of the previous iteration. (d) EMI-FGSM accumulates the average gradient of the sampled data points in the direction of $\bar{g}_{t-1}$ for update, where $\bar{g}_{t-1}$ is the average gradient of the previous iteration.

Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) [18] integrates Nesterov's accelerated gradient (NAG) [24] into the iterative attack method to further improve the transferability of adversarial examples:

$$\tilde{x}_t^{adv} = x_t^{adv} + \alpha \cdot \mu \cdot g_{t-1}, \quad g_t = \mu \cdot g_{t-1} + \frac{\nabla_{\tilde{x}_t^{adv}} J_f(\tilde{x}_t^{adv}, y)}{\|\nabla_{\tilde{x}_t^{adv}} J_f(\tilde{x}_t^{adv}, y)\|_1}, \quad x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_t).$$

3.2 Pre-gradient Guided Momentum based Attack

As shown in Figure 2 (a), MI-FGSM [5] accumulates the gradient of each iteration to stabilize the update direction and escape from poor local maxima, and achieves higher transferability than I-FGSM [14]. As depicted in Figure 2 (b), NI-FGSM *looks ahead* by accumulating the gradient after adding momentum to the current data point so as to converge faster and achieve higher transferability [18].

The performance improvement of NI-FGSM over MI-FGSM is mainly due to the *looking ahead* property of the Nesterov's accelerated gradient. We observe that NI-FGSM adopts the accumulated momentum in MI-FGSM to *look ahead*, which is designed to obtain more stable direction by considering the history gradient. This inspires us to study a new problem: *Although the direction of accumulated momentum helps craft more transferable adversaries, is it the optimal direction for looking ahead?*

To explore the direction of *looking ahead*, we propose a variation of NI-FGSM, called the Pre-gradient guided momentum Iterative FGSM (PI-FGSM), which *looks ahead* by the gradient of the previous iteration. Specifically, as shown in Figure 2 (c), PI-FGSM accumulates the gradient of data point obtained by adding the previous gradient to the current data point at each iteration. The update procedure can be summarized as:

$$\tilde{x}_t^{adv} = x_t^{adv} + \alpha \cdot \tilde{g}_{t-1}, \quad \tilde{g}_t = \nabla_{\tilde{x}_t^{adv}} J_f(\tilde{x}_t^{adv}, y),$$
$$g_t = \mu \cdot g_{t-1} + \frac{\tilde{g}_t}{\|\tilde{g}_t\|_1}, \quad x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_t),$$

where $\tilde{g}_{t-1}$ denotes the gradient of the previous iteration. Instead of considering all the history gradient as in NI-FGSM, PI-FGSM *looks ahead* guided by the local gradient information and achieves better attack performance, as demonstrated in Sec. 4.2.

3.3 Enhanced Momentum based Attack

We continue to investigate the family of momentum based attacks and observe that at each iteration, MI-FGSM [6], NI-FGSM [18] and our PI-FGSM accumulate the gradient of different data points, and they all exhibit higher transferability than I-FGSM [14] that only adopts the gradient of the current data point for update. This indicates that the accumulated gradient is helpful for crafting highly transferable adversaries. Since the accumulation of gradient of these methods are on different data points, this inspires us another question: *At each iteration, could we further improve the attack transferability by accumulating the gradients of multiple data points around the current data point for the iterative gradient-based attacks?*

To address this question, we enhance the momentum by not only memorizing all the past gradients during the iterative process, but also accumulating the gradients of multiple sampled examples in the vicinity of the current data point. Considering the performance improvement of PI-FGSM, to help sample more useful data points for the gradient calculation, we sample multiple data points along the direction used in PI-FGSM, *i.e.* the gradient direction of the previous iteration. Specifically, as shown in Figure 2 (d), we calculate the gradient of the t-th iteration as follows:

$$\bar{x}_t^{adv}[i] = x_t^{adv} + c_i \cdot \bar{g}_{t-1} \tag{1}$$

$$\bar{g}_t = \frac{1}{N}\sum_{i=1}^{N} \nabla_{\bar{x}_t^{adv}[i]} J_f(\bar{x}_t^{adv}[i], y), \tag{2}$$

where N is the sampling number, $\bar{g}_{t-1}$ is the gradient calculated at the previous iteration and c_i is the i-th coefficient sampled in interval $[-\eta, \eta]$. In our experiments, we adopt linear sampling, which samples N linearly spaced data points in the interval $[-\eta, \eta]$ and also try other sampling methods as shown in Sec. 4.5. We denote such accumulated gradient as the enhanced momentum.

Note that the proposed enhanced momentum is generally applicable to any iterative gradient-based attacks, such as I-FGSM [14], PGD [22], and the ensemble-model attack [19]. Here we incorporate the enhanced momentum into I-FGSM, denoted as Enhanced Momentum I-FGSM (EMI-FGSM), to craft highly transferable adversarial examples. The update procedure can be summarized as:

$$g_t = \mu \cdot g_{t-1} + \frac{\bar{g}_t}{\|\bar{g}_t\|_1}, \tag{3}$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \mathrm{sign}(g_t). \tag{4}$$

where $\bar{g}_t$ is calculated by Eq. (2). The algorithm is summarized in Algorithm 1.

4 Experiments

In this section, we provide the experimental setup, report comparisons of gradient-based attacks on four normally trained models and comparisons when integrated with input transformations and ensemble-model attack, as well as results of attacking seven advanced defenses. We further provide ablation studies for the sampling method and hyper-parameters. We also give discussions on other possible variant methods in Appendix B. Code is available at https://github.com/JHL-HUST/EMI.

4.1 Experimental Setup

Dataset. Similar to [18, 35, 39], we randomly choose 1,000 images from the ILSVRC 2012 validation set [28]. All these images are resized to $299 \times 299 \times 3$ beforehand.

Algorithm 1 EMI-FGSM.

Input: A classifier f and loss function J_f. A benign example x and its ground-truth label y.
Input: The maximum perturbation ε, number of iteration T and decay factor μ. The bound η for the sampling interval and sampling number N.
Output: An adversarial example $x^{adv} \in \mathcal{B}_\varepsilon(x)$.

$\alpha = \varepsilon/T$; $g_0 = 0$; $\bar{g}_0 = 0$; $x_1^{adv} = x$.
for $t = 1 \to T$ **do**:
  Sample N coefficients $c_i \in [-\eta, \eta]$ for Eq. (1).
  Calculate the average gradient $\bar{g}_t$ of N sampled data points by Eq. (2).
  Update the enhanced momentum g_t by Eq. (3).
  Update x_{t+1}^{adv} by Eq. (4).
end for
return $x^{adv} = x_{T+1}^{adv}$.

Baselines. We compare our method with six gradient-based attack methods including FGSM [11], I-FGSM [14], PGD [22], CW [2], MI-FGSM [5] and NI-FGSM [18]. We also integrate our method into the ensemble-model attack [5, 19], input transformations [6, 18, 39], and Patch-wise attack [8] to show the performance improvement of our method over these baselines.

Models. Four normally trained models, *i.e.* Inception-v3 (Inc-v3) [32], Inception-v4 (Inc-v4), Inception-Resnet-v2 (IncRes-v2) [33], Resnet-v2-101 (Res-101) [12], as well as three ensemble adversarially trained models, *i.e.* ens3-adv-Inception-v3 (Inc-v3$_{ens3}$), ens4-Inception-v3 (Inc-v3$_{ens4}$), ens-adv-Inception-ResNet-v2 (IncRes-v2$_{ens}$) [34], are considered. Without ambiguity, we simply call the three ensemble adversarially trained models as *adversarially trained models*. Moreover, to show the efficacy of our methods, we also incorporate seven advanced defense methods, including the top-3 submission in the NIPS 2017 defense competition, *i.e.* high-level representation guided denoiser (HGD, rank-1) [17], input transformation through random resizing and padding (R&P, rank-2) [38], NIPS-r3 (rank-3) [1], randomized smoothing (RS) [3] and adversarially randomized smoothing (ARS) [29] for certified defense, feature distillation (FD) [20] and bit depth reduction (Bit-Red) [40]. Here we do not consider some SOTA adversarial training methods, *e.g.* PGD-AT [22], TRADES [41], which only validates the effectiveness on CIFAR-10 or CIFAR-100 datasets.

Attack Settings. We follow the settings in [5] with the maximum perturbation of $\varepsilon = 16/255$, pixels normalized into $[0,1]$ and the number of iteration $T = 10$. For the momentum term, we set the decay factor $\mu = 1$ [5]. For DIM, we set the transformation probability to 0.5 and the input x is first randomly resized to an $r \times r \times 3$ image with $r \in [299, 330)$, and then padded to $330 \times 330 \times 3$ [39]. For TIM, we adopt Gaussian kernel with size 7×7 [6]. For SIM, the number of scale copy is set to $m = 5$ [18]. For EMI-FGSM, we set the number of examples $N = 11$, the sampling interval bound $\eta = 7$, and adopt the linear sampling.

4.2 Comparison with Gradient-based Attacks

We first craft adversaries by various gradient-based attacks in single-model and ensemble-model setting respectively, and report the attack success rates, which are the misclassification rates of the corresponding models using adversarial examples as the inputs.

Single-model Setting. The results for adversaries crafted on Inc-v3 are depicted in Ta-

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

Attack	Inc-v3*	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
FGSM	67.3	25.7	26.0	24.5	10.2	10.4	4.5
I-FGSM	**100.0**	20.3	18.5	16.1	4.6	5.2	2.5
PGD	98.3	17.8	15.5	11.9	5.8	5.9	3.3
CW	**100.0**	19.6	15.7	13.3	4.1	5.2	2.4
MI-FGSM	**100.0**	44.5	42.0	36.3	13.4	13.7	6.5
NI-FGSM	**100.0**	51.9	50.4	41.0	13.4	13.2	5.7
PI-FGSM (**Ours**)	**100.0**	60.2	59.1	49.0	14.9	14.6	6.5
EMI-FGSM (**Ours**)	**100.0**	**72.7**	**69.9**	**59.5**	**20.3**	**19.9**	**10.9**

Table 1: Attack success rates (%) against seven baseline models in single-model setting. The adversaries are crafted on Inc-v3.* indicates the white-box model being attacked.

Attack	Inc-v3*	Inc-v4*	IncRes-v2*	Res-101*	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
FGSM	64.8	49.3	43.9	68.8	15.8	15.1	8.9
I-FGSM	99.9	98.6	95.6	99.8	19.1	16.8	10.4
PGD	**100.0**	99.7	99.4	99.9	11.4	11.0	6.7
CW	**100.0**	**99.8**	98.8	**100.0**	15.4	14.4	9.5
MI-FGSM	99.9	98.7	95.0	99.9	39.7	35.5	23.8
NI-FGSM	**100.0**	**99.8**	99.2	99.9	41.2	34.9	22.9
PI-FGSM (**Ours**)	**100.0**	99.2	98.5	99.9	52.5	45.3	29.6
EMI-FGSM (**Ours**)	**100.0**	**99.8**	**99.8**	**100.0**	**69.0**	**62.0**	**43.0**

Table 2: Attack success rates (%) against seven baseline models in ensemble-model setting. The adversaries are crafted on ensemble models, *i.e*. Inc-v3, Inc-v4, IncRes-v2 and Res-101.

ble 1 and the results on other three models are shown in Appendix A. All the attacks achieve 100% attack success rates in white-box setting except for FGSM. For black-box attacks, some attacks (*e.g*. I-FGSM, PGD, CW) that have demonstrated high effectiveness in white-box setting, exhibit low transferability when evaluated on other models. On the contrary, the transferability of PI-FGSM is much higher (8-9%) on normally trained models, and is considerably higher (0.8-1.5%) on adversarially trained models than MI-FGSM and NI-FGSM. With the enhanced momentum, EMI-FGSM exhibits much higher transferability on normally trained models (10.5-12.5%) and adversarially trained models (4.4-5.4%) than PI-FGSM, and outperforms NI-FGSM with a clear margin of 11.1% on average.

Ensemble-model Setting. As in [5], we implement the attacks in ensemble-model setting by fusing the logit outputs of four normally trained models, *i.e*. Inc-v3, Inc-v3, IncRes-v2 and Res-101, with equal ensemble weights. As shown in Table 2, I-FGSM, PGD and CW attack still show lower transferability than other attacks. In contrast, PI-FGSM exhibits better attack success rates than I-FGSM and MI-FGSM in white-box setting and achieves much higher transferability on three adversarially trained models. This validates our first concern that due to considering too much history gradient, the accumulated momentum adopted by NI-FGSM provides imprecise direction compared with PI-FGSM, which is not optimal for looking ahead. Moreover, EMI-FGSM achieves the best results in both white-box and black-box setting and outperforms the powerful baseline NI-FGSM by a large margin of more than 20%, which demonstrates the high effectiveness of enhanced momentum.

4.3 Integrated with Input Transformations or Patch-wise attack

We further incorporate EMI-FSGM with various input transformations, *i.e*. DIM, TIM, SIM, and the combination of three input transformations, dubbed DTS for abbreviation and patch-wise attack (PIM), in single-model and ensemble-model setting respectively. For fairness, all the transformations are integrated into MI-FGSM as baselines [6, 39]. We integrate PIM into I-FGSM, MI-FGSM and our EMI-FGSM, termed PIM, MI-PIM and EMI-PIM, respectively.

Attack	Inc-v3*	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
DIM	99.0	64.6	60.9	52.1	18.3	17.7	9.5
EMI-DIM (**Ours**)	**99.1**	**83.5**	**78.0**	**70.6**	**27.8**	**26.0**	**13.4**
TIM	**100.0**	47.0	44.5	40.5	24.3	22.0	13.2
EMI-TIM (**Ours**)	**100.0**	**79.4**	**76.3**	**67.2**	**44.3**	**40.8**	**26.2**
SIM	**100.0**	70.3	68.0	62.4	32.4	30.8	17.2
EMI-SIM (**Ours**)	**100.0**	**91.9**	**90.0**	**85.4**	**45.2**	**41.8**	**23.8**
DTS	98.9	83.1	80.7	75.8	65.2	62.7	46.0
EMI-DTS (**Ours**)	**99.6**	**94.1**	**92.6**	**89.4**	**78.9**	**75.3**	**60.4**
PIM	**100.0**	43.4	32.3	36.2	33.2	34.9	24.3
MI-PIM	**100.0**	50.7	**46.6**	**43.9**	18.3	19.8	10.5
EMI-PIM (**Ours**)	99.7	**51.7**	40.2	43.5	**42.6**	**44.8**	**33.4**

Table 3: Attack success rates (%) against seven baseline models in single-model setting. The adversaries are crafted on Inc-v3.

Attack	Inc-v3*	Inc-v4*	IncRes-v2*	Res-101*	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
DIM	99.4	97.4	94.7	**99.8**	56.3	50.7	36.4
EMI-DIM (**Ours**)	**99.9**	**99.6**	**99.7**	99.7	**77.0**	**70.1**	**50.3**
TIM	99.8	98.0	95.0	99.9	61.3	56.7	47.8
EMI-TIM (**Ours**)	**100.0**	**100.0**	**99.7**	**100.0**	**89.0**	**83.9**	**78.2**
SIM	99.9	99.3	98.5	**100.0**	78.5	74.4	60.4
EMI-SIM (**Ours**)	**100.0**	**100.0**	**100.0**	**100.0**	**90.1**	**87.3**	**74.2**
DTS	99.6	98.9	97.9	99.7	92.1	90.2	86.6
EMI-DTS (**Ours**)	**100.0**	**99.9**	**100.0**	**100.0**	**97.4**	**96.1**	**94.1**
PIM	**100.0**	99.9	99.9	99.7	60.2	60.3	45.6
MI-PIM	**100.0**	**100.0**	**100.0**	**99.9**	42.7	41.1	29.3
EMI-PIM (**Ours**)	**100.0**	99.9	99.8	99.8	**84.6**	**86.2**	**80.6**

Table 4: Attack success rates (%) against seven baseline models in ensemble-model setting. The adversaries are crafted on ensemble models, *i.e.* Inc-v3, Inc-v4, IncRes-v2 and Res-101.

Single-model Setting. The results for adversaries generated on Inc-v3 are summarized in Tabel 3. We can observe that EMI can significantly boost the transferability on each of the transformation-based attack methods. In general, the EMI based attacks consistently outperform the baseline attacks by 3.9% $\sim$ 32.4%. Even for white-box setting, EMI further promotes the attack success rates of the baseline attacks. For instance, EMI-DTS outperforms DTS by 0.7% against Inc-v3. For PIM, we find that MI-PIM boosts the transferability on clean models but degrades the transferability on adversarially trained models. In contrast, EMI-PIM enhance the transferability of PIM on both clean models and adversarial trained models. The results for adversaries crafted on other three normally-trained models are consistent with that generated on Inc-v3, as shown in Appendix A.

Ensemble-model Setting. As in Sec. 4.2, we also evaluate the attacks in ensemble-model setting and the results are summarized in Table 4. EMI based method remarkably improves the attack success rates across all experiments over the baseline attacks. In particular, the final combination of EMI-DTS has achieved the attack success rates of over 94.1% for black-box attacks against the three adversarially trained models. Such intriguing results convincingly demonstrate the success on the combination of EMI-FGSM, input transformations and ensemble-model attack for improving the attack transferability.

4.4 Attacking Advanced Defense Models

With the remarkable improvement on the baselines, we further evaluate EMI-FGSM on seven advanced defenses with various input transformations in ensemble-model setting to show its high efficacy. We test the defenses using the adversaries crafted in Sec. 4.3.

Attack	Sampling Method	Inc-v3*	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
EMI-FGSM	Linear	**100.0**	74.4	**71.9**	60.5	**21.2**	**19.5**	**9.9**
	Uniform	**100.0**	**74.7**	71.5	**61.2**	18.9	18.8	8.8
	Gaussian	**100.0**	73.0	70.4	60.0	20.2	19.0	**9.9**
EMI-DTS	Linear	**99.6**	94.5	**92.8**	**90.2**	**78.8**	**76.0**	**60.3**
	Uniform	99.5	92.9	91.9	87.9	77.1	71.5	57.0
	Gaussian	99.5	**94.8**	92.6	89.7	78.5	74.3	58.7

Table 5: Attack success rates (%) of EMI-FGSM and EMI-DTS against the seven baseline models with various sampling methods. The adversarial examples are crafted on Inc-v3.

The results of EMI-FGSM with three input transformations are shown in Figure 3(a)-3(c). EMI-FGSM remarkably improves the transferability of three input transformations on all defense models. On average, the performance is improved by 14.8%, 24.5% and 11.8% respectively. We also integrate the combination of three input transformations into EMI-FGSM as in [18] to further improve the transferability. As shown in Figure 3(d), EMI-DTS achieves an average attack success rate of 86.6%, boosting SOTA methods by a clear margin of 7.8%. Given that the adversaries are crafted on the ensemble models without any defense mechanisms but with such high attack performance, it identifies the inefficiency of existing defenses and indicates that they are far from being deployed in real-world applications.

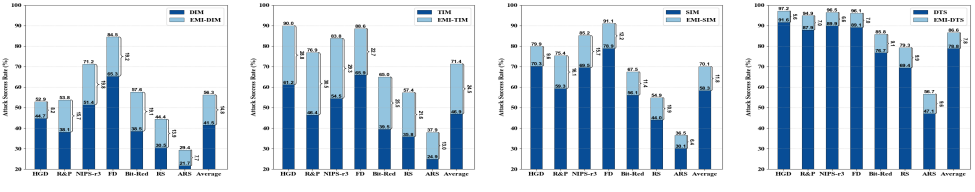
4.5 Ablation Study

To gain more insights on the performance improvement by enhanced momentum based methods, we conduct ablation studies to explore the impact of sampling method and hyper-parameters for sampling interval η and sampling number N, respectively. To simplify the analysis, we only consider the transferability of adversaries crafted on Inc-v3 by vanilla EMI-FGSM and EMI-DTS. The default setting adopts linear sampling, $\eta = 7$ and $N = 11$.

On sampling distribution. We try EMI-FGSM and EMI-DTS with three types of sampling methods, *i.e.*, linear sampling, uniform sampling and Gaussian sampling. Linear sampling samples N linearly spaced data points in the interval. Uniform and Gaussian sampling sample N data points in the interval by uniform and Gaussian distribution, respectively. As shown in Table 5, the three sampling methods achieve similar attack performance. In general, linear sampling exhibits slightly higher results, thus we adopt linear sampling in experiments.

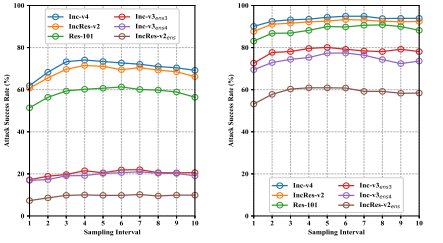
On sampling interval. To validate the impact of sampling interval η, we try different values of η from 1 to 10 and the results are summarized in Figure 4. For all the values of η, the white-box attack success rate is 100%. The transferability increases when $\eta \leq 3$ for both EMI-FGSM and EMI-DTS. For $4 \leq \eta \leq 7$, the attacks exhibit similar transferability and the performance decays slightly when $\eta > 7$. Thus we adopt $\eta = 7$ in experiments.

On sampling number. We explore the impact of the sampling number N, as shown in Figure 5. The white-box attack success rate for various values of N is 100%. When $N = 1$,



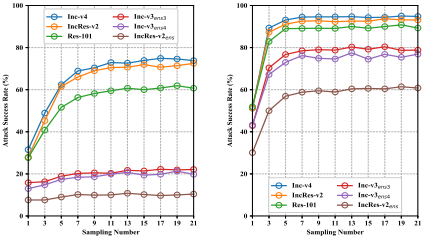
(a) DIM *vs.* EMI-DIM (b) TIM *vs.* EMI-TIM (c) SIM *vs.* EMI-SIM (d) DTS *vs.* EMI-DTS

Figure 3: Attack success rates (%) against seven advanced defenses. The adversaries are crafted on ensemble models, *i.e.* Inc-v3, Inc-v4, IncRes-v2, Res-101. (Zoom in for details.)



(a) EMI-FGSM (b) EMI-DTS

Figure 4: Attack success rates (%) on the other six models with adversarial examples generated by EMI-FGSM and EMI-DTS on Inc-v3 for various **sampling interval**.



(a) EMI-FGSM (b) EMI-DTS

Figure 5: Attack success rates (%) on the other six models with adversarial examples generated by EMI-FGSM and EMI-DTS on Inc-v3 for various **sampling number**.

Attack	Inc-v3*	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
MI-FGSM	**100.0**	39.3	36.6	31.5	8.9	8.6	3.3
EMI-FGSM (**Ours**)	**100.0**	**72.7**	**69.9**	**59.5**	**20.3**	**19.9**	**10.9**

Table 6: Attack success rates (%) of MI-FGSM (T=110) and EMI-FGSM (T=10, N=11) against the seven baseline models under the same computational cost.

EMI-FGSM degrades to MI-FGSM and exhibits the lowest transferability. When we increase the value of N, the transferability increases rapidly before $N = 11$ for EMI-FGSM and $N = 7$ for EMI-DTS. When $N > 11$, increasing N can still bring small performance improvement for EMI-FGSM. However, the bigger the value of N, the higher the computational cost. To balance the performance gain and the cost, we set $N = 11$ in experiments.

4.6 Discussion on Computational Cost

The computational cost for gradient-based attacks mainly depends on the forward and backward propagation for the gradient calculation, which is related to the number of gradient calculation at each iteration N and the total number of iterations T. Under the same number of iterations, EMI-FGSM needs N times number of gradient calculation, leading to N times overhead. To further validate the effectiveness under the same cost, we set $T = 110$ for MI-FGSM while $T = 10$ and $N = 11$ for our EMI-FGSM. As shown in Table 6, with larger number of iterations, MI-FGSM tends to overfit the target model and achieves even lower transferability than MI-FGSM with 10 iterations. This indicates larger computational cost does not guarantee better transferability and shows the superiority of our method.

5 Conclusion

Inspired by momentum based attacks, we propose an enhanced momentum that accumulates the gradient of each iteration as well as the gradients of the sampled data points in the gradient direction of previous iteration. Empirical evaluations on ImageNet dataset demonstrate that our enhanced momentum can significantly improve the attack success rates in white-box and black-box settings. Our EMI-DTS, integrated with input transformations in ensemble-model setting, could achieve an average black-box attack success rate of over 94%, showing very high transferability. Our work also indicates that existing defenses are far from being deployed in real-world applications and stronger robust deep learning models are needed.

Acknowledgements

This work is supported by National Natural Science Foundation (62076105) and Microsft Research Asia Collaborative Research Fund (99245180).

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning (ICML)*, 2018.

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

[3] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning (ICML)*, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019.

[5] 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 (CVPR)*, pages 9185–9193, 2018.

[6] 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 (CVPR)*, pages 4312–4321, 2019.

[7] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1625–1634, 2018.

[8] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *European Conference on Computer Vision (ECCV)*, 2020.

[9] Lianli Gao, Qilong Zhang, Jingkuan Song, and Heng Tao Shen. Patch-wise++ perturbation for adversarial targeted attacks. *arXiv preprint arXiv:2012.15503*, 2020.

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 2015.

[12] 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 (CVPR)*, pages 770–778, 2016.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[14] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *International Conference on Learning Representations (ICLR), Workshop Track Proceedings*, 2017.

[15] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 641–649, 2020.

[16] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. *International Conference on Machine Learning (ICML)*, 2019.

[17] 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 (CVPR)*, pages 1778–1787, 2018.

[18] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2020.

[19] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *International Conference on Learning Representations (ICLR)*, 2017.

[20] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868, 2019.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3431–3440, 2015.

[22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018.

[23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[24] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o(1/k^2). *Doklady AN USSR*, 269:543–547, 1983.

[25] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[26] 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.

[27] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[28] 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 (IJCV)*, 115(3):211–252, 2015.

[29] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11292–11303, 2019.

[30] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM Sigsac Conference on Computer and Communications Security*, pages 1528–1540, 2016.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*, 2014.

[32] 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 (CVPR)*, pages 2818–2826, 2016.

[33] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[34] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations (ICLR)*, 2018.

[35] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.

[36] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. *International Conference on Computer Vision (ICCV)*, 2021.

[37] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[38] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *International Conference on Learning Representations (ICLR)*, 2018.

[39] 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 (CVPR)*, pages 2730–2739, 2019.

[40] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Network and Distributed System Security Symposium (NDSS)*, 2018.

[41] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *International Conference on Machine Learning (ICML)*, 2019.

[42] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. On success and simplicity: A second look at transferable targeted attacks. *arXiv preprint arXiv:2012.11207*, 2020.