Improving Adversarial Transferability with Spatial Momentum

Guoqiu Wang, Xingxing Wei†, Huanqian Yan

Beijing Key Laboratory of Digital Media, Beihang University, Beijing, China
{wangguoqiu, xxwei, yanhq}@buaa.edu.cn

Abstract

Deep Neural Networks (DNN) are vulnerable to adversarial examples. Although many adversarial attack methods achieve satisfactory attack success rates under the white-box setting, they usually show poor transferability when attacking other DNN models. Momentum-based attack (MI-FGSM) is one effective method to improve transferability. It integrates the momentum term into the iterative process, which can stabilize the update directions by adding the gradients’ temporal correlation for each pixel. We argue that only this temporal momentum is not enough, the gradients from the spatial domain within an image, i.e. gradients from the context pixels centered on the target pixel are also important to the stabilization. For that, in this paper, we propose a novel method named Spatial Momentum Iterative FGSM Attack (SMI-FGSM), which introduces the mechanism of momentum accumulation from temporal domain to spatial domain by considering the context gradient information from different regions within the image. SMI-FGSM is then integrated with MI-FGSM to simultaneously stabilize the gradients’ update direction from both the temporal and spatial domain. The final method is called SM\textsuperscript{2}I-FGSM. Extensive experiments are conducted on the ImageNet dataset and results show that SM\textsuperscript{2}I-FGSM indeed further enhances the transferability. It achieves the best transferability success rate for multiple mainstream undefended and defended models, which outperforms the state-of-the-art methods by a large margin.

Introduction

Though deep neural networks (DNNs) have achieved state-of-the-art performance on various vision tasks, including image classification (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014), object detection (Girshick 2015; Ren et al. 2015; Redmon et al. 2016) and semantic segmentation (Chen et al. 2017a; Long, Shelhamer, and Darrell 2015), they are vulnerable to adversarial examples (Goodfellow, Shlens, and Szegedy 2015; Szegedy et al. 2013) which are crafted by adding imperceptible perturbations to clean images, making models output wrong predictions expected by attackers. The existence of adversarial examples has raised concerns in security-sensitive applications, e.g., self-driving cars (Liu et al. 2019a), face recognition (Xiao et al. 2021) and video monitoring (Li et al. 2019).

In the past years, many methods have been proposed to generate adversarial examples, such as fast gradient sign method (Goodfellow, Shlens, and Szegedy 2015) and its iterative variants (Kurakin, Goodfellow, and Bengio 2016), projected gradient descent method (Madry et al. 2017), C&W attack (Carlini and Wagner 2017) and so on. They all conduct attacks in the white-box setting, which utilize the detailed information of the threat models. In addition, some works show that adversarial examples have transferability (Liu et al. 2016; Papernot et al. 2017), which means the adversarial examples crafted for one DNN model can successfully attack other DNN models to a certain extent. The existence of transferability makes adversarial examples practical to the real-world applications because attackers do not need to know the information of the target models, and thus introduces a series of serious security issues (Xiao et al. 2021; Liu et al. 2019a).

However, the above white-box attack methods usually have poor transferability. Recently, many methods have been proposed to address this issue, such as momentum-based iterative attack (Dong et al. 2018; Lin et al. 2019), variance
tuning iterative gradient-based method (Wang and He 2021), diverse inputs attack (Xie et al. 2019), translation-invariant attack (Dong et al. 2019), scale-invariant attack (Lin et al. 2019), and multi-model ensemble attack (Liu et al. 2016). However, there is still a gap between transferability attack success rates and the practical demand, which motivates us to design a more effective method to further improve adversarial transferability when attacking various DNN models.

Among the above methods to improve transferability, the momentum-based iterative attack (MI-FGSM) (Dong et al. 2018) show good performance, and have many variants (Lin et al. 2019; Wang and He 2021; Wang et al. 2021). They integrate the momentum term into the iterative process, which can stabilize the update directions during the iterations by adding the gradients’ temporal correlation to obtain better perturbations. We argue that only this temporal momentum is not enough, the gradients from the spatial domain within an image, i.e., gradients from the context pixels centered on the target pixel are also important to the stabilization. For that, in this paper, we propose a novel method named Spatial Momentum Iterative FGSM (SMI-FGSM) attack, which introduces the mechanism of momentum accumulation from temporal domain to spatial domain by considering the context gradient information from different regions within the image. By smoothing the gradient in this way, we can get a stable gradient used for generating adversarial examples, which achieve good generalization to various DNN models. SMI-FGSM is then integrated with the previous MI-FGSM to construct SM²I-FGSM, and thus simultaneously enhances the transferability from both the temporal and spatial domain. This process is illustrated in Figure 1.

The main contributions can be summarized as follows:

• We show that, in addition to the previous temporal gradient momentum, the gradient momentum coming from the spatial domain is also useful to enhance transferability. Moreover, we show that multiple random transformations will lead to an effective spatial momentum.
• According to the above idea, we propose a novel method called Spatial Momentum Iterative (SMI-FGSM) attack to improve the transferability. SMI-FGSM is then integrated with the MI-FGSM to construct the final SM²I-FGSM, which further enhances the transferability from both the temporal and spatial domain.
• Extensive experimental results show that the proposed method could remarkably improve the attack transferability in both mainstream undefended and defended models.

Related Work
Adversarial Attacks
The study of adversarial attack is considered in developing robust models (Madry et al. 2017). It can be roughly categorized into two types, white-box attacks, and black-box attacks. The deep models are fully exposed to the adversary in the white-box attack setting, as their structure and parameters. Whereas in the black-box attack setting, the adversary only has little or no knowledge of the target model. Hence, black-box attacks are more practical in real-world scenarios.

Black-box attacks mainly include query-based attacks (Chen et al. 2017b; Tu et al. 2019) and transfer-based attacks (Liu et al. 2016; Dong et al. 2019). Query-based attacks focus on estimating the gradients of the target model through interaction with the target model. However, these methods usually require a large number of queries, which is unrealistic in real-world applications. Transfer-based attacks, which are more practical and have been studied extensively, generate adversarial examples by using a white-box attack method on a source model (or source models in ensemble attack) to fool the target model. Here, we focus on improving the transfer-based black-box attacks in this paper.

Defend against Adversarial Attacks
Several methods have been proposed to defend adversarial examples, which generally fall into two axes. The first one is termed adversarial training (Goodfellow, Shlens, and Szegedy 2015; Kurakin, Goodfellow, and Bengio 2017b), where adversarial examples are generated to train models to withstand adversarial examples, which generally fall into two axes. The first one is termed adversarial training (Goodfellow, Shlens, and Szegedy 2015; Kurakin, Goodfellow, and Bengio 2017b), which injects adversarial examples into the training procedure, the adversarially trained models learn to resist the perturbations in the gradient direction of the loss function. Ensembl adversarial training (Tramer et al. 2018) augments the training data with the adversarial samples produced not only from the model being trained but also from other holdout models. Therefore, the ensemble adversarially trained models are more robust against adversarial attacks.

The second line of defenses is carried out by input transformation to purify the adversarial examples. They preprocess the inputs to cleanse adversarial perturbations without reducing the classification accuracy on clean images (Wu et al. 2021). The advanced defenses of this kind include applying random resizing and padding (R&P) (Xie et al. 2018), a high-level representation denoiser (HGD) (Liao et al. 2018), JPEG compression (JPEG) (Guo et al. 2018), feature distillation (FD) (Liu et al. 2019b), feature squeezing method: bit reduction (BIT) (Xu, Evans, and Qi 2018) and a neural representation purifier (NRP) (Naseer et al. 2020). In this paper, we exploit these state-of-the-art defenses to evaluate the effectiveness of the proposed attack.

Methodology
In this section, we give an overview of the family of gradient-based adversarial attacks first. Then we introduce the motivation and provide detailed descriptions of the proposed SMI-FGSM and SM²I-FGSM. We also analyze the differences between temporal momentum attacks and spatial momentum attacks.

Gradient-based Adversarial Attack Methods
Given a classification network \( f_θ \) parameterized by \( θ \), let \( (x, y) \) denote the clean image and its corresponding ground-truth label, the goal of adversarial attack is to find an example \( x^{adv} \) which is in the vicinity of \( x \) but misclassified by the network. In most cases, we use the \( L_p \) norm to limit the adversarial perturbations below a threshold \( ϵ \), where \( p \) could be 0, 2, ∞. This can be expressed as

\[
f_θ(x^{adv}) \neq y, \quad \text{s.t.} \quad \|x^{adv} - x\|_p \leq ϵ \quad (1)
\]
Fast Gradient Sign Method (FGSM) \cite{goodfellow2015explaining} generates an adversarial example $x_{adv}$ by performing one-step update as

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y))$$ \hfill (2)

where $\nabla_x J$ is the gradient of the loss function $J(\cdot)$ with respect to $x$ and cross-entropy loss is often used. $\text{sign}(\cdot)$ is the sign function to limit perturbations conform to the $L_\infty$ norm bound.

The Iterative version of FGSM (I-FGSM) \cite{kurakin2016adversarial} iteratively applies fast gradient sign method multiple times with a small step size $\alpha$, which can be expressed as

$$x_{adv}^{t+1} = x_{adv}^t + \alpha \cdot \text{sign}(\nabla_x J(x_{adv}^t, y))$$ \hfill (3)

where $x_{adv}^0 = x$. I-FGSM induces much more powerful white-box attacks than FGSM but less transferability \cite{kurakin2016adversarial}.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM) \cite{dong2018boosting} boosts the transferability of adversarial examples by integrating a momentum term into the iterative attack method to stabilize the update directions. The update procedure is

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_{adv}^t, y)}{\left\| \nabla_x J(x_{adv}^t, y) \right\|_1}$$ \hfill (4)

$$x_{adv}^{t+1} = x_{adv}^t + \alpha \cdot \text{sign}(g_{t+1})$$ \hfill (5)

where $g_t$ is the accumulated gradient and $\mu$ is the decay factor which is often set to 1.0. MI-FGSM updates the moment gradient $g_{t+1}$ by Eq. (4) and then updates $x_{adv}^{t+1}$ by Eq. (5).

Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) \cite{lin2019variance} adapts nesterov accelerated gradient into the iterative attacks so as to effectively look ahead and improves the transferability of adversarial examples. NI-FGSM substitutes $x_{adv}^t$ in Eq. (4) with $x_{adv}^t + \alpha \cdot \mu \cdot g_t$.

Variance Tuning Momentum-based Iterative Method (VMI-FGSM) \cite{wang2021variance} further consider the gradient variance of the previous iteration to tune the current gradient so as to stabilize the update direction. It substitutes Eq. (4) by

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_{adv}^t, y) + v_t}{\left\| \nabla_x J(x_{adv}^t, y) + v_t \right\|_1}$$ \hfill (6)

where $v_{t+1} = \frac{1}{n} \sum_{i=1}^{n} \nabla_x J(x_i, y) - \nabla_x J(x_{adv}^t, y)$, $x_i = x_{adv}^t + r_i$; and $r_i$ is the random noise within a certain range.

Diverse Inputs (DI) attack \cite{xie2019improving} applies random transformations to the input images at each iteration to create hard and diverse input patterns, which brings randomness to the adversarial perturbations and improves the generalization of adversarial examples efficiently. The transformation includes random resizing and padding with a given probability.

Translation-Invariant (TI) attack \cite{dong2019translation} shifts the image within small magnitude and approximately evenly translates image with a pre-defined kernel matrix. The result adversarial example is less sensitive to the discriminative region of the white-box model being attacked and has a higher probability to fool another model, especially for black-box models with defense mechanisms.

Scale-Invariant (SI) attack \cite{lin2019variance} introduces the scale-invariant property of deep learning models and optimizes the adversarial perturbations over the scale copies by a factor $1/2^2$ of the input image to enhance the adversarial transferability.

Spatial Momentum Iterative Attack

In this section, we will introduce the motivation and spatial momentum iterative attack in detail. We show the algorithm and its experimental results under the constraint of $L_\infty$ norm. This method can also be used in $L_2$ norm.

The original momentum method accelerates the gradient descent algorithms across iterations by accumulating the velocity vector in the gradient direction of the loss function \cite{polyak1964some}. MI-FGSM \cite{dong2018boosting} introduces the idea of momentum into adversarial attack and gets a big promotion. It integrates the momentum term into the iterative process, which can be seen as adding temporal correlation (see Eq. (4)) for the gradient that is used to update perturbations compared to I-FGSM \cite{kurakin2016adversarial} and VMI-FGSM \cite{dong2018boosting} and can stabilize the update directions during the iterations. It motivates us that momentum accumulation mechanism not only can be based on a temporal domain like \cite{dong2018boosting}, but also in a spatial domain through comprehensively considering the context pixels centered on the target pixel within the image.

As Eq. (6) shows, I-FGSM simply updates the perturbation with the gradient from image $x_{adv}^t$, which only considers the current pixel, while ignores its context pixels. In order to stabilize the direction of updating, SMI-FGSM calculate the gradients by convolving the gradient of the untranslated image with a pre-defined kernel matrix. The result adversarial example is less sensitive to the discriminative region of the white-box model being attacked and has a higher probability to fool another model, especially for black-box models with defense mechanisms.

Figure 2: The stabilization of gradient during each iteration. We use cosine similarity of gradients during iteration to measure the stabilization. We can see that SMI-FGSM achieves better stabilization than I-FGSM and SM²I-FGSM achieves the best stabilization among the four methods. The results are averaged over 1000 images.
Figure 3: Illustrations of temporal momentum-based attacks, spatial momentum-based attacks and their variants. There are some symbols such as \( g_t \), \( \alpha \), \( \mu \), \( v_t \), etc. Please refer to subsection Gradient-based Adversarial Attack Methods for detail. (a) Temporal momentum-based iterative attack. (b) Spatial momentum-based iterative attack. (c) Temporal and spatial momentum-based iterative attack SM²I-FGSM.

Algorithm 1: SM²I-FGSM Algorithm

**Input:** Classifier \( f(\cdot) \) with loss function \( J(\cdot) \), clean image \( x \) and ground-truth label \( y \),\( n \) denotes the number of transformation, decay factor \( \beta \), iterations \( T \), max perturbation \( \epsilon \).

**Output:** adversarial example \( x_{adv} \)

1. \( \alpha = \epsilon / T \);
2. \( x_{adv}^0 = x \), \( g_0 = 0 \);
3. for \( t = 0 \) to \( T - 1 \) do
4. for \( i = 0 \) to \( n - 1 \) do
5. Get the gradient of \( i \)-th transformed image by 
   \[ G_i = \nabla_x J(H_i(x_{adv}^t), y) \]
6. end for
7. Obtain spatial momentum gradient \( g_{s+1}^t \) by 
   \[ g_{s+1}^t = \frac{1}{n} \sum \lambda_i G_i \]
8. Update \( g_{t+1} \) by 
   \[ g_{t+1} = \beta \cdot g_t + \frac{g_{s+1}^t}{\|g_{s+1}^t\|_2} \]
9. Update \( x_{adv}^{t+1} \) by Eq. (7)
10. end for
11. return \( x_{adv} = x_{adv}^T \)

The Difference with Existing Attacks

As shown in Figure 3, temporal momentum-based methods stabilize the direction by using historical gradients. The basic method is MI-FGSM, which updates the current gradient using the previous gradient (illustrated by the solid line). NI-FGSM improves it by Nesterov accelerated gradient and VMI-FGSM boosts it by considering the gradient variance by adding various noises. These two methods use the previous gradient two times, which play different roles (see the solid line and dotted line in Figure 3). However, they do not consider the spatial domain information, which is the same important as temporal domain information. The proposed SMI-FGSM considers spatial domain information by randomly transforming the image several times during
Figure 4: The transferability attack success rates (%) against four models with adversarial examples generated by SM\textsuperscript{2}I-FGSM on Inc-v3 and Inc-v4 when varying $n$.

each iteration (illustrated by the yellow boxes). TI-FGSM also smoothes gradients in the spatial domain, which uses a simple pre-defined convolution kernel. However, its context scope is limited, and experiments show that its performance is also limited (see Table 1). In our method, by combining temporal and spatial momentum, SM\textsuperscript{2}I-FGSM can further stabilize the direction and achieves better performance.

Experiments

In this section, we first introduce experimental settings, followed by parameter tuning and ablation study for the proposed methods. The comparisons with state-of-the-art attacks are finally reported and discussed.

Experimental Settings

We randomly sample 1000 images of different categories from the ILSVRC 2012 validation set (Russakovsky et al. 2015) as in (Dong et al. 2018; Wang and He 2021). We also ensure that all of the selected images can be correctly classified by every model exploited in this work.

To evaluate our approach and compare with other mainstream methods, we test attack performance in three normally trained models, including Inception-v3 (Inc-v3) (Szegedy et al. 2016), Inception-v4 (Inc-v4) and Inception-Resnet-v2 (IRes-v2) (Szegedy et al. 2017), and two adversarially trained models, i.e., ens-adv-Inception-v3 (Inc-v3\textsubscript{ens4}) and ens-adv-InceptionResNet-v2 (IRes-v2\textsubscript{ens4}) (Tramer et al. 2018). The official models are used here. In addition, three input transformation based defense strategies, including FD (Liu et al. 2019b), BIT (Xu, Evans, and Qi 2018), and NRP (Naseer et al. 2020), are used to purify adversarial images. After input transformations, the purified images are fed to Inc-v3\textsubscript{ens4} to give the final prediction. In this way, we have eight target models to test the performance of different attacks.

For the settings of hyper-parameters, we follow the setting in (Dong et al. 2018) with the maximum perturbation $\epsilon = 16$ among all experiments with pixel values in $[0, 255]$, the number of iteration $T = 10$, and step size $\alpha = 1.6$. For MI-FGSM, we set $\mu = 1.0$ as recommend in (Dong et al. 2018). For the transformation function $H(\cdot)$, the random size is in $[270, 299]$ and then padding it to 299. We adopt the Gaussian kernel with kernel size 5 $\times$ 5 for translation-invariant. And for our proposed SM\textsuperscript{2}I-FGSM, we set $n = 12$ and $\beta = 1.0$. We use Attack Success Rates (ASR), which refers to the percentage of all images that can be misclassified by the target model. The bigger value, the better.

Impact of Hyper-parameter $n$

In SM\textsuperscript{2}I-FGSM, the number of transformations $n$ plays a key role in improving the transferability attack success rates. When $n$ is set to 1, SM\textsuperscript{2}I-FGSM will degenerate to MI-FGSM. Therefore, we practiced a series of experiments to examine the effect of $n$. We attack Inc-v3 and Inc-v4 by SM\textsuperscript{2}I-FGSM with different $n$ values, which range from 2 to 20 with a granularity 2, and the results are shown in Figure 4. As shown, the increase of transferability attack success rates is rapid at the beginning and then leveled off when $n$ exceeds 12. Considering attack ability and computational complexity, $n$ is set to 12 in our experiments.
Table 1: The attack success rates (%) of I-FGSM, TI-FGSM, MI-FGSM, SMI-FGSM, and SM\textsuperscript{2}I-FGSM under single-model setting, * indicates the white-box model being attacked. The best results are marked in bold. We evaluate the attacks on normally trained models (i.e., Inc-v3, Inc-v4, and IRes-v2), adversarially trained models (i.e., Inc-v3\text\_ens and IRes-v2\text\_ens), and input transformation defense strategies (i.e., FD, BIT and NBR).

| Model  | Attack      | Inc-v3 | Inc-v4 | IRes-v2 | Inc-v3\text\_ens | IRes-v2\text\_ens | FD | BIT | NBR | Average |
|--------|-------------|--------|--------|---------|------------------|------------------|----|-----|-----|---------|
| Inc-v3 | I-FGSM      | 99.8*  | 25.6   | 21.9    | 13.2             | 5.4              | 12.8| 9.5 | 6.7 | 13.6    |
|        | TI-FGSM     | 99.9*  | 28.3   | 24.5    | 15.0             | 6.6              | 14.9| 12.8| 9.5 | 15.3    |
|        | MI-FGSM     | 100.0* | 49.3   | 47.9    | 30.7             | 18.7             | 28.4| 20.1| 16.9| 30.3    |
|        | SMI-FGSM    | 100.0* | 59.6   | 53.8    | 34.2             | 20.0             | 29.5| 25.7| 21.2| 34.8    |
|        | SM\textsuperscript{2}I-FGSM | 99.8* | 78.5   | 76.1    | 61.6             | 48.0             | 58.9| 47.5| 49.1| 59.9    |
| Inc-v4 | I-FGSM      | 18.7   | 99.3*  | 11.1    | 7.1              | 3.2              | 8.7 | 6.5 | 7.4 | 9.0     |
|        | TI-FGSM     | 23.4   | 99.4*  | 13.5    | 7.8              | 5.3              | 11.0| 10.4| 9.7 | 11.6    |
|        | MI-FGSM     | 43.1   | 99.3*  | 34.1    | 21.7             | 12.8             | 21.1| 17.6| 15.4| 23.7    |
|        | SMI-FGSM    | 47.7   | 99.2*  | 36.7    | 17.5             | 11.4             | 23.2| 16.8| 17.3| 24.4    |
|        | SM\textsuperscript{2}I-FGSM | 76.0 | 99.5*  | 67.8    | 46.5             | 38.3             | 39.8| 33.6| 29.2| 47.3    |
| IRes-v2| I-FGSM      | 18.1   | 13.1   | 98.6*   | 7.7              | 4.6              | 8.1 | 4.3 | 5.6 | 8.8     |
|        | TI-FGSM     | 22.7   | 18.9   | 98.4*   | 12.3             | 8.5              | 11.8| 9.0 | 12.3| 13.7    |
|        | MI-FGSM     | 43.6   | 36.2   | 98.8*   | 22.2             | 18.7             | 19.9| 15.0| 16.4| 24.6    |
|        | SMI-FGSM    | 45.5   | 38.9   | 97.5*   | 21.8             | 16.3             | 18.5| 16.1| 15.6| 24.6    |
|        | SM\textsuperscript{2}I-FGSM | 73.1 | 69.3   | 97.5*   | 52.3             | 49.8             | 42.8| 36.5| 40.1| 51.9    |

Table 2: The attack success rates (%) of various NI-FGSM, VMI-FGSM, and SM\textsuperscript{2}I-FGSM under single-model setting, * indicates the white-box model being attacked. The best results are marked in bold.

| Attack     | Inc-v3 | Inc-v4 | IRes-v2 | Inc-v3\text\_ens | IRes-v2\text\_ens | FD | BIT | NBR | Average |
|------------|--------|--------|---------|------------------|------------------|----|-----|-----|---------|
| Inc-v3     | NI-FGSM | 100.0* | 53.9   | 54.3             | 34.0             | 23.3| 29.4| 24.3| 20.7    | 34.3    |
|            | VMI-FGSM | 100.0* | 68.9   | 66.7             | 47.8             | 41.9| 46.6| 38.2| 42.8    | 50.4    |
|            | SM\textsuperscript{2}I-FGSM | 99.8* | 78.5   | 76.1             | 61.6             | 48.0| 58.9| 47.5| 49.1    | 59.9    |
| Inc-v4     | NI-FGSM | 47.1   | 99.8*  | 37.8             | 21.3             | 12.9| 20.9| 19.7| 18.3    | 25.5    |
|            | VMI-FGSM | 70.6   | 99.6*  | 61.5             | 40.2             | 35.7| 33.1| 28.7| 25.3    | 42.1    |
|            | SM\textsuperscript{2}I-FGSM | 76.0 | 99.5*  | 67.8             | 46.5             | 38.3| 39.8| 33.6| 29.2    | 47.3    |
| IRes-v2    | NI-FGSM | 45.8   | 39.5   | 97.0*            | 22.7             | 19.5| 21.8| 18.9| 19.3    | 26.8    |
|            | VMI-FGSM | 68.9   | 66.2   | 97.2*            | 47.5             | 42.7| 33.5| 29.8| 31.7    | 45.8    |
|            | SM\textsuperscript{2}I-FGSM | 73.1 | 69.3   | 97.5*            | 52.3             | 49.8| 42.8| 36.5| 40.1    | 51.9    |

Ablation Study

We first perform adversarial attacks using I-FGSM and SMI-FGSM under a single-model setting. The results are reported in Table 1. The attacked DNN models are listed on rows, and the test DNN models are listed on columns. It is obvious that SMI-FGSM is strong as I-FGSM when attacking white-box models, they all have nearly 100% success rate. It can be seen that the attack based on spatial momentum has significantly improved adversarial attack transferability. For example, when we generate adversarial examples using Inc-v3 as a white-box model, SMI-FGSM achieves success rates of 20.0% on IRes-v2\text\_ens, while I-FGSM, TI-FGSM, and MI-FGSM achieve success rates of 5.4%, 6.6%, and 18.7% on IRes-v2\text\_ens respectively. Through comparison and statistics, the adversarial attack transferability of our proposed method is ahead about 20% than all baseline methods on average, which directly demonstrates the effectiveness of the proposed method. It’s worth noting, SM\textsuperscript{2}I-FGSM outperforms MI-FGSM by a large margin. This reveals the importance of spatial information for improving transferability.

Comparisons with State-of-the-art Attacks

We use NI-FGSM and VMI-FGSM as comparison algorithms, they are all improved versions of MI-FGSM and can generate adversarial images with much higher transferability. We compare the performance of MI-FGSM and its improved versions (i.e., NI-FGSM, VMI-FGSM, SM\textsuperscript{2}I-FGSM) in Table 2. SM\textsuperscript{2}I-FGSM outperforms the others by a large margin and it is model-agnostic. Particularly, if we generate adversarial images on Inc-v3, SM\textsuperscript{2}I-FGSM achieves an average success rate of 59.9%, while the state-of-the-art methods achieve 34.3%, and 50.4% respectively. We show several adversarial images generated by MI-FGSM, NI-FGSM, VMI-FGSM, and SM\textsuperscript{2}I-FGSM in Figure 5. It can be seen that SM\textsuperscript{2}I-FGSM generates visually similar adversarial perturbation as others, which demonstrates the superiority of the proposed attack method.

Performances Combined with Other Methods

Diverse inputs (DI), translation-invariant (TI), and scale-invariant (SI) can further improve the attack success rates individually based on I-FGSM and MI-FGSM. Lin et al.
Table 3: The attack success rates (%) of MI-FGSM-DTS, NI-FGSM-DTS, VMI-FGSM-DTS, and SM²I-FGSM-DTS under a single-model setting. ∗ indicates the white-box model being attacked. The best results are marked in bold.

| Model          | Attack         | Inc-v3 | Inc-v4 | IRes-v2 | Inc-v3* | IRes-v2* | FD | BIT | NBR | Average |
|----------------|----------------|--------|--------|---------|---------|---------|-----|-----|-----|---------|
| Inc-v3         | MI-FGSM-DTS    | 99.9*  | 82.9*  | 80.5*   | 70.0*   | 57.7*   | 70.4| 43.8| 48.3| 64.8    |
|                | NI-FGSM-DTS    | 99.8*  | 83.7*  | 82.3*   | 71.2*   | 57.9*   | 71.0| 44.9| 45.5| 65.2    |
|                | VMI-FGSM-DTS   | 99.6*  | 83.5*  | 82.6*   | 77.4*   | 63.5*   | 74.6| 50.9| 56.4| 69.8    |
|                | SM²I-FGSM-DTS  | 99.8*  | 87.1*  | 86.9*   | 80.7*   | 67.2*   | 79.0| 57.2| 58.1| 81.9    |

| Inc-v4         | MI-FGSM-DTS    | 84.0   | 99.7   | 79.8*   | 64.7*   | 53.0*   | 65.3| 36.1| 32.5| 59.4    |
|                | NI-FGSM-DTS    | 86.5   | 99.7*  | 80.3*   | 63.0*   | 51.9*   | 65.5| 36.9| 30.9| 59.3    |
|                | VMI-FGSM-DTS   | 89.3   | 99.8*  | 84.1*   | 73.1*   | 61.8*   | 70.4| 43.1| 39.5| 65.9    |
|                | SM²I-FGSM-DTS  | 92.0   | 100.0* | 88.2*   | 75.3*   | 63.6*   | 77.6| 67.6| 41.3| 80.5    |

| IRes-v2        | MI-FGSM-DTS    | 77.0   | 73.7*  | 97.3*   | 60.3*   | 57.4*   | 67.0| 38.9| 42.7| 59.6    |
|                | NI-FGSM-DTS    | 77.9   | 74.3*  | 97.1*   | 60.7*   | 57.0*   | 67.5| 40.8| 44.6| 60.4    |
|                | VMI-FGSM-DTS   | 79.2   | 77.8*  | 97.0*   | 66.3*   | 62.7*   | 70.9| 48.8| 50.4| 65.2    |
|                | SM²I-FGSM-DTS  | 80.3   | 79.0   | 98.1*   | 67.9*   | 64.8*   | 76.9| 57.2| 53.1| 73.5    |

(Lin et al. 2019) have shown that the combination of them, which is called DTS in this paper, could help the gradient-based attacks achieve great transferability. We combine DTS with MI-FGSM, NI-FGSM, VMI-FGSM and SM²I-FGSM as MI-FGSM-DTS, NI-FGSM-DTS, VMI-FGSM-DTS and SM²I-FGSM-DTS. The results are reported in Table 3. From the table, we can observe that SM²I-FGSM-DTS achieves an average transferability success rate of 81.9% when crafting adversarial images on Inc-v3 model. Compared to the baseline method MI-FGSM-DTS, which achieves an average transferability success rate of 64.8%, this is a significant improvement and shows that our method has good scalability and can be combined with existing methods to further improve the success rate of transfer-based black-box attacks.

**Ensemble-based Attacks**

Related work (Liu et al. 2016) has shown that the transferability success rate can be greatly improved by generating adversarial examples using multiple models. We also test its effect on our method. There are three ensemble methods, i.e., ensemble in logits, ensemble in predictions, and ensemble in the loss function. Here we fuse the logit outputs of different models, which is the most common ensemble method among the three (Dong et al. 2018). In this subsection, we perform ensemble-based attacks by averaging the logit outputs of the models Inc-v3, Inc-v4, and IRes-v2. And the results are recorded in Table 4. SM²I-FGSM-DTS achieves an average attack success rate up to 90.8% on five defense models. It is worth noting that when the transferability attack success rate exceeds 90% in defense strategies, our method is still 2% higher than the most advanced attack, which shows the effectiveness of the proposed method and indicates the vulnerability of current defense mechanisms.

**Conclusion**

In this paper, we proposed a spatial momentum method for improving the transferability of adversarial examples, which introduce the mechanism of momentum accumulation from the temporal domain to the spatial domain. And it can be well integrated with existing attack strategies to further improve the adversarial transferability. Extensive experimental results show that the proposed methods could remarkably improve the attack transferability in both excellent undefended and defended models under the single-model or multi-model settings. By comparing with the most advanced attacks, it further demonstrates the effectiveness of the proposed method. Specifically, our attack algorithm SM²I-FGSM-DTS can achieve a 90.8% transferability attack success rate on the most advanced defense strategies on average, which indicates the vulnerability of current defense mechanisms and inspire us to develop more robust models.
References

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (sp), 39–57. IEEE.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4): 834–848.

Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; and Hsieh, C.-J. 2017b. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Proceedings of the 10th ACM workshop on artificial intelligence and security, 15–26.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In Proceedings of the IEEE conference on computer vision and pattern recognition, 9185–9193.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4312–4321.

Girshick, R. 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, 1440–1448.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Guo, C.; Rana, M.; Cissé, M.; and van der Maaten, L. 2018. Countering Adversarial Images using Input Transformations. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25: 1097–1105.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2017a. Adversarial examples in the physical world. arXiv:1607.02533.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2016. Adversarial examples in the physical world. CoRR, abs/1607.02533.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2017b. Adversarial Machine Learning at Scale. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Li, S.; Neupane, A.; Paul, S.; Song, C.; Krishnamurthy, S. V.; Roy-Chowdhury, A. K.; and Swami, A. 2019. Stealthy Adversarial Perturbations Against Real-Time Video Classification Systems. In NDSS.

Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; and Zhu, J. 2018. Defense against adversarial attacks using high-level representation guided denoiser. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1778–1787.

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2019. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In International Conference on Learning Representations.

Liu, A.; Liu, X.; Fan, J.; Ma, Y.; Zhang, A.; Xie, H.; and Tao, D. 2019a. Perceptual-sensitive gan for generating adversarial patches. In Proceedings of the AAAI conference on artificial intelligence, 1028–1035.
Wang, X.; and He, K. 2021. Enhancing the Transferability of Adversarial Attacks Through Variance Tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1924–1933.

Wang, X.; Lin, J.; Hu, H.; Wang, J.; and He, K. 2021. Boosting Adversarial Transferability through Enhanced Momentum. arXiv preprint arXiv:2103.10609.

Wu, W.; Su, Y.; Lyu, M. R.; and King, I. 2021. Improving the Transferability of Adversarial Samples With Adversarial Transformations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9024–9033.

Xiao, Z.; Gao, X.; Fu, C.; Dong, Y.; Gao, W.; Zhang, X.; Zhou, J.; and Zhu, J. 2021. Improving Transferability of Adversarial Patches on Face Recognition With Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11845–11854.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. L. 2018. Mitigating Adversarial Effects Through Randomization. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2730–2739.

Xu, W.; Evans, D.; and Qi, Y. 2018. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In 25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018. The Internet Society.