Boosting the Transferability of Adversarial Examples with Translation Transformation

Zheming Li*, Hengwei Zhang†, Xiaohu Liu and Jindong Wang
PLA SSF Information Engineering University, Zhengzhou 450001, China
*These authors contributed equally to this work
†Corresponding Author: Hengwei Zhang. Email: zhw11qd@aliyun.com

Abstract. Although the adversarial examples have achieved an incredible white-box attack rate, they tend to show poor transferability in black-box attacks. Date augmentation is considered to be an effective means of enhancing the adversarial examples transferability. To this end, based on translation transformation we propose a new method to generate more mobile adversarial examples to attack the advanced defense model. By optimizing the original image, the input diversity is improved and the transferability of the adversarial examples generated by further training is enhanced. This method can also be combined with the attack method based on gradient, and the new method can make the attack success rate higher. Experiments on ImageNet data sets show that the proposed method is superior to gradient based methods such as MI-FGSM in black box attack, while maintaining a high success rate of white box attack. We hope that our proposed method of attack can serve as a benchmark for assessing the robustness of networks to opponents and the effectiveness of the different methods of defence.

1. Introduction
Convolutional neural network has advanced performance in image classification and recognition, and has been widely used. However, they are vulnerable and cannot distinguish the content of the image properly, because they are vulnerable to adversarial examples, which are elaborately made by adding human imperceptible interference to benign images. Confrontational examples bring uncertainty to image recognition and security risks to many application scenarios, such as face recognition [1], automatic driving [2], which has aroused widespread attention and promoted the research on the attack and defense.

According to the degree of mastering the model information, the attack methods can be divided into white-box attack and black-box attack: Under the white-box setting, the attacker has a clear understanding of the network structure and parameters, and the existing attack methods have achieved a high success rate. However, under the black-box condition, the transferability of the generated adversarial examples is poor. This means that adversarial examples created for one known model cannot attack another unknown model with great success. In order to solve this problem, many methods have been proposed, such as Momentum Iterative Fast Gradient Sign Method [3], which belongs to gradient-based attacks, Ensemble-Model Attacks and Input Transformation Based Attacks [4]. Nevertheless, there is still a big difference in the success rate between white-box attacks and black-box attacks.

The generation process of adversarial instance can be compared to the training process of neural network model. Correspondingly, the transitivity of adversarial instance can be compared to the generalization ability of the model [5]. Input transformation can be likened to data augmentation in
model training, which increases the number of images in training set through scaling, cropping, rotation and other operations. The application of this method to the generation of adversarial examples can produce good results, effectively improve the transitivity of adversarial examples, and produce better attack effect on unknown models.

For this reason, we propose the Translation Transformation Attack Method (TTM) to improve the portability of adversarial examples. We randomly load the translation operation of the image onto the original image in each iteration. Specifically, it translates images horizontally and vertically, and whether the image is translated and the pixel size of the translation is random. Figure 1 shows the comparison of the image generated by TTM method with the original image and the image generated by other attack methods.

![Figure 1. Images generated by different methods of generating adversarial examples. The generation methods are as follows: the fast gradient sign method (FGSM) [6], the Iterative method [7], the Momentum Iterative Fast Gradient Sign Method [3] and the proposed Translation Transformation Momentum Iterative Fast Gradient Sign Method (TT-MI-FGSM) for the Inception v3 [8] model.

2. Related Work

Deep neural networks (DNNs) are vulnerable to attacks against adversarial examples. From the perspective of application, adversarial examples can be used to attack the trained model, so as to measure the robustness of the model; It can also be used as training data to further train the model to improve the recognition ability of unknown images and the ability to resist malicious attacks.

2.1. Generating Adversarial Examples

The goal of generating adversarial examples is to maximize the interference factor on the image and the loss function of the classification algorithm on the premise that the image interference does not exceed the visible limit of human eyes. According to this idea, the generation method of adversarial examples can be roughly divided into two kinds: one generation method based on optimized loss function, the other based on optimized input image. Szegedy et al. [9] first pointed out the vulnerability of convolutional neural network, and proposed the method of L-BFGS attack model network. Later, Goodfellow et al. [10] proposed a fast gradient signal method to generate adversarial examples by a one-step method. Kurakin et al. [11] introduced the idea of iteration and proposed i-fgsm, which improved the success rate of white box attack. However, due to over fitting, the success rate of black box attack decreased. In order to improve the transferability of counterwork cases, Dong et al. [3] introduced the idea of momentum in the process of generating counterwork samples, and proposed a method based on momentum, which improved the success rate of black box attack. Later, the method of data augmentation was applied in the generation of adversarial examples. Xie et al. [12] proposed the method
of DI2-FGSM, which greatly improved the success rate of the black box attack by improving the training
diversity of adversarial examples. Dong et al. [13] proposed the TI-FGSM, which improves the
transferability of adversarial examples by calculating the gradient of a group of images.

2.2. Defending Against Adversarial Examples

Several defense methods are proposed to strengthen the model to prevent it from being attacked. At first,
some people [14-15] proposed that the adversarial examples could be mixed into the normal training
data to improve the robustness of the training model. Tramèr et al. [16] proposed ensemble adversarial
training, in other words, it enhanced the training data by enhancing the perturbations from the other
models. Prakash et al. [17] proposed a framework combining pixel migration and soft wavelet denoising
to resist adversarial examples. [18] used generation model to purify hostile images and moved them
back to the distribution of clean images.

3. Methodology

In this part, we briefly introduce the main contents of fast gradient sign methods, and introduce the
methods we proposed in detail. Let $x$ be the input of the whole adversarial examples generation system,
$y^{\text{true}}$ denote the corresponding ground-truth label, meanwhile the $\theta$ are the parameters of the model,
$L(x, y^{\text{true}}; \theta)$ is the loss function of neural network, which is usually the cross entropy loss function in
practical application. For the process of generating adversarial examples, we maximize $L(x, y^{\text{true}}; \theta)$ to
generate a $x_{\text{adv}} = x + r$ that cannot distinguish from $x$ by to fool the model. In general, we use the infinite
norm to measure the perceptibility of the disturbance $r$, that is, the infinite norm of $r$ should be less than $\epsilon$ . Therefore, the goal of adversarial example generation is to achieve the following mathematical goals
through optimization:

$$
\arg\max_{x_{\text{adv}}} L(x_{\text{adv}}, y^{\text{true}}; \theta), \quad s.t. \|r\|_{\infty} \leq \epsilon .
$$

(1)

3.1. Family of Fast Gradient Sign Methods

——Fast Gradient Sign Method: This method is the initial version of this method and the simplest of
the algorithms. The algorithm maximizes the loss function in a single step by calculating the gradient of
the input image $x$ loss function. This method is expressed by the following mathematical formula:

$$
x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y^{\text{true}}; \theta))
$$

(2)

Where $x_{\text{adv}}$ is the image of the generated adversarial examples, and $\epsilon$ is the size of the added
disturbance. The method achieves a high white box success rate, but the transferability of attacks needs
to be improved.

——Iterative Fast Gradient Sign Method: The loss function is iteratively updated along the current
gradient direction through several small steps $\alpha$, which can be expressed as:

$$
x_{\text{adv}}^{0} = x ,
$$

$$
x_{n+1}^{\text{adv}} = \text{Clip}_x \{ x_{n}^{\text{adv}} + \alpha \cdot \text{sign}(\nabla_x L(x_{n}^{\text{adv}}, y^{\text{true}}; \theta)) \}
$$

(3)

Where $\alpha$ is the step size in each iteration process and $\alpha = \epsilon / T$, in which $T$ is the total number of
iterations. Research shows that I-FGSM can produce more powerful white box attack than FGSM, but
the cost is worse portability

——Momentum Iterative Fast Gradient Sign Method: The momentum factor is added in the attack
process to make the update direction of the loss function tend to be stable, so that the loss function
exceeds the local maximum. The process of updating is similar to that of I-FGSM, as shown below:

$$
x_{0}^{\text{adv}} = x ,
$$

$$
g_0 = 0 ,
$$
\[ g_{n+1} = \mu \cdot g_n + \frac{\nabla L(x_{n, \text{adv}}, y\text{true}; \theta)}{\| \nabla L(x_{n, \text{adv}}, y\text{true}; \theta) \|} , \]  
\[ x_{n, \text{adv}}^{n+1} = \text{Clip}_{\varepsilon} \{ x_{n, \text{adv}}^n + \alpha \cdot \text{sign}(g_{n+1}) \} , \]
g \_ n is the value of gradient accumulated during the \( t \)-th, and \( \mu \) is the delay factor of \( g \). Compared with the 1-FGSM, This method greatly improves the performance of black box attack while ensuring the success rate of white-box attack is basically unchanged.

—Diverse Input Method: The method used the data augmentation method to the generation process, applied random transformation to the input, and calculated the gradient of the transformed image. The random scaling and zero filling method used in this paper can also be used as a supplementary method to combine with other gradient based attack methods to form a new attack method, which further improves the transferability of adversarial examples.

### 3.2. Translation Transformation Attack Method

The generation process of adversarial examples have the same idea to the training process of network model. Meanwhile the transferability of well-made adversarial examples and the portability of training model are similar. Under the guidance of this idea, the translation transformation of graphs can also be used to the generation process. We propose the translation transformation attack method, and the basic process is as follows:

\[ x_{0, \text{adv}} = x \]
\[ x_{n+1, \text{adv}} = \text{Clip}_{\varepsilon} \{ x_{n, \text{adv}} + \alpha \cdot \text{sign}(\nabla L(T(x_{n, \text{adv}}; p), y\text{true}; \theta)) \} \]

the function is:

\[ T(x_{n, \text{adv}}; p) = \begin{cases} T(x_{n, \text{adv}}), & p \\ x_{n, \text{adv}}, & p - 1 \end{cases} \]

Where \( T(\cdot) \) is the random translation transformation of the image, that is, after the image is input into the generation system, it will be transformed by the random pixel size in the direction of up, down, left and right with the probability of \( p \). Then, the gradients of the transformed image are calculated, and the images with the maximum loss function are obtained along the gradient direction, which are the adversarial examples.

### 3.3. Attack algorithms

TTM uses data augmentation to prevent overfitting in generating adversarial examples. TTM can be formed in combination with the MI-FGSM to optimize from two dimensions, to produce better transferability adversarial examples. This method can be called the Translation Transformation Momentum Iterative Fast Gradient Sign Method (TT-MI-FGSM).

| TT-MI-FGSM |
|----------------|
| **Input:** Original input image \( x \); label file \( y\text{true} \); classifier model file \( f \) with loss function \( L \); artificial disturbance added in the generation process \( \varepsilon \); maximum number of iterations with noise \( T \); decay factor \( \mu \) relative to the last iteration. |
| **Output:** An adversarial example \( x_{\text{adv}} \) |
| 1: \( \alpha = \varepsilon / T \) |
| 2: \( x_{0, \text{adv}} = x \); \( g_0 = 0 \) |
| 3: for \( t = 0 \) to \( T - 1 \) do |
| 4: Get \( x_{\text{adv}} \) by \( y_{\text{adv}} = T(x_{\text{adv}}; p) \) > apply translation transformation to the inputs image with the probability \( p \) |
| 5: Get the gradients by \( \nabla L(x_{\text{adv}}; y\text{true}; \theta) \) |
6: Update \( g_{n+1} \) by \( g_{n+1} = \mu \cdot g_n + \nabla_x L(\chi_{adv}^{n+1}, y_{adv}; \theta) \)

7: Update \( x_{n+1}^{adv} \) by Eq. (5)

8: return \( x_{adv} = x_T \)

4. Experiments

4.1. Experimental setup

Dataset. We selected 1000 correctly classified images from the ImageNet dataset, and all the images were pre-processed to a size of 299*299*3.

Models. In the experiment, we use a total of seven network models, they are four normally networks, i.e., Inception-v3 (Inc-v3) [8], Inception-v4 (Inc-v4), Inception-Resnet-v2 (IncRes-v2), (Res-101) and three confrontation training networks ens3-adv-inception -v3 (Inc-v3ens3), ENS4-ADV-Inception - V3 (INC-V3ENS4) and ENS-ADV-ADV-Inception - ResNet-V2 (INGS-V2ENS). The adversary examples are generated on the normal training network model, and then tested on seven network models.

Baselines. We combine our method with I-FGSM [11] and MI-FGSM [3] to form a new combination method, and input the generated adversarial examples into the network classifier to evaluate the improvement of TTM compared with these baseline methods.

Parameter setting. In the experiment, we need to set the necessary parameters, we set the maximum disturbance \( \epsilon = 16 \), step size \( \alpha = 1.6 \), and number of iterations \( T = 10 \). The decay factor is used to reflect the real situation in the generation process, and we set it \( \mu = 1.0 \). P is used to measure the probability of translation transformation, \( p \) is set to 0.5 in our methods.

4.2. Attacking a single network

In this section, under the premise of single model, we will use our method to attack the model, and compare the success rate with the baseline attack method. We put the results in Table 1, we can see clearly that our extended method is always better than the baseline attack, and the success rate is close to 100% under the white box setting. This shows that TT-MI-FGSM can be used as a powerful method to attack single model effectively, and the migration is improved effectively.

Table 1: In the single model setting, the success rate (%) of the seven models against the attack. The confrontational examples are compiled with I-FGSM, MI-FGSM and TT-MI-FGSM on inc-v3, inc-v4, incres-v2 and res-101 respectively. White box attack is indicated by marking *.

| Model   | Attack       | Inc-v3 | Inc-v4 | IncRes-v2 | Res-101 | Inc-v3ens3 | Inc-v3ens4 | IncRes-v2ens |
|---------|--------------|--------|--------|-----------|---------|------------|------------|-------------|
| Inc-v3  | I-FGSM       | 99.9*  | 22.7   | 20.3      | 18.2    | 7.1        | 7.6        | 4.2         |
|         | MI-FGSM      | 99.9*  | 48.2   | 47.1      | 39.9    | 15.3       | 14.2       | 7.2         |
|         | TT-MI-FGSM (Ours) | 99.0* | 70     | 66.6      | 58.7    | 21.2       | 21.2       | 9.1         |
| Inc-v4  | I-FGSM       | 37.8   | 99.9*  | 26.1      | 21.9    | 8.6        | 8.0        | 5.1         |
|         | MI-FGSM      | 63.8   | 99.9*  | 53.7      | 47.7    | 19.7       | 16.8       | 9.4         |
|         | TT-MI-FGSM (Ours) | 79    | 99.4*  | 69.8      | 62.2    | 25.6       | 22.5       | 13.1        |
| IncRes-v2 | I-FGSM       | 37.3   | 31.9   | 99.6*     | 25.9    | 8.8        | 7.6        | 4.9         |
|         | MI-FGSM      | 68.6   | 61.9   | 99.6*     | 52.5    | 25.1       | 20.2       | 14.6        |
|         | TT-MI-FGSM (Ours) | 80.1  | 76.8   | 98.9*     | 67.8    | 35.9       | 30.1       | 19.9        |
| Res-101 | I-FGSM       | 27.8   | 23.4   | 21.4      | 98.2*   | 9.4        | 7.9        | 5.7         |
|         | MI-FGSM      | 52.4   | 48.3   | 45.6      | 98.2*   | 22.3       | 18.7       | 11.8        |
|         | TT-MI-FGSM (Ours) | 72.7  | 70.5   | 65.6      | 97.8*   | 33.6       | 29.5       | 18.2        |
4.3. Attacking an ensemble of networks

Ensemble network model is considered to have good defense performance. Experiments on the ensemble model can better prove the effectiveness of our method, so we attack multiple models at the same time, and the experimental results further prove the performance of our method. We consider to show the performance of our methods by attacking multiple models (including Inc-v3, Inc-v4, IncRes-v2 and Res-101) simultaneously. Specifically, we first set the set of models, and set the weight of each training model to equal weight. Then we use I-FGSM, MI-FGSM and TMI-FGSM to attack, and we can get the attack effect of different attack methods.

As shown in Table 2, we will experiment on the classifier with the adversarial examples generated by the ensemble model. Through data comparison, our method improves the attack success rate. By calculating the average attack success rate, we can see that the average performance of our method is better than the baseline attack by 5.3% under the black box setting. The results show that the robustness of these advanced adversarial training models under TT-MI-FGSM black box attack is very small.

Table 2: Attack success rate (%) against seven models in multi model environment. Adversarial examples are crafted for a collection of inc-v3, inc-v4, incres-v2, and res-152. White box attack is indicated by marking *.

| Attack                  | Inc-v3* | Inc-v4* | IncRes-v2* | Res-101* | Inc-v3ens3 | Inc-v3ens4 | IncRes-v2ens3 | IncRes-v2ens4 | Average |
|-------------------------|---------|---------|------------|----------|------------|------------|----------------|----------------|---------|
| I-FGSM                  | 99.7    | 96.2    | 91.9       | 86.6     | 18.8       | 15.9       | 9.3            | 59.8           |
| MI-FGSM                 | 99.8    | 97.7    | 95.1       | 91.2     | 38.4       | 36.0       | 22.4           | 68.7           |
| TT-MI-FGSM (Ours)       | 99.1    | 96.9    | 95.2       | 92.0     | 54.1       | 49.3       | 31.8           | 74.0           |

5. Conclusions

In this paper, we propose a new adversarial examples generation method named Translation Transformation Momentum Iterative Fast Gradient Sign Method (TT-MI-FGSM). Specifically, through the random translation of the original image, the diversity of the input image is realized, and the number of samples is increased, so as to improve the migration of adversarial instances. After that, we carry out experiments on ImageNet data set, and the results show that compared with other existing gradient based attack methods, this method can significantly improve the black box attack, and maintain a higher success rate of white box attack. In addition, tests on ensemble model show that our method can outperform MI-FGSM by 12.7%, indicating the insufficiency of current defense techniques. It is hoped that the proposed attack method can train more robust networks and become a standard for evaluating robustness.

Acknowledgments

This project was supported by the National Key Research and Development Program of China under Grant no. 2017YFB0801900.

References

[1] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. (2016) Accessorize to a crime: Real and stealthyattacks on state-of-the-art face recognition. In: Proceedings of the 2016 acm sigsac conference on computer and communications security, pages 1528–1540,
[2] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. (2018) Robust Physical-World Attacks on Deep Learning Models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[3] Y. Dong et al. (2018) Boosting adversarial attacks with momentum, in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit., Salt Lake City, UT, USA, pp. 9185–9193.
[4] Xiaosen Wang, Xuanran He, Jingdong Wang and Kun He, (2021) Admix: Enhancing the Transferability of Adversarial Attacks, arXiv:2102.00436, [Online]. Available: http://arxiv.org/abs/2102.00436.
[5] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. (2020) Nesterov accelerated gradient and scale invariance for adversarial attacks. In International Conference on Learning Representations (ICLR).
[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. (2015) Explaining and harnessing adversarial examples. In ICLR.
[7] A. Kurakin, I. Goodfellow, and S. Bengio, (2016) Adversarial examples in the physical world, arXiv: 1607.02533. [Online]. Available: https://arxiv.org/abs/1607.02533.
[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, JonShlens, and Zbigniew Wojna. (2016) Rethinking the inception architecture for computer vision. In CVPR, 2016.
[9] C. Szegedy et al., (2013) Intriguing properties of neural networks, arXiv: 1312.6199. [Online]. Available: https://arxiv.org/abs/1312.6199.
[10] Goodfellow, I.J., Shlens, J., Szegedy, C. (2015) Explaining and harnessing adversarial examples. In: International Conference on Learning Representations.
[11] Kurakin, A., Goodfellow, I., Bengio, S. (2017) Adversarial machine learning at scale. In: International Conference on Learning Representations.
[12] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. (2019) Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2730–2739.
[13] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. (2019) Evading defenses to transferable adversarial examples by translation-invariant attacks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4312–4321.
[14] Goodfellow, I.J., Shlens, J., Szegedy, C. (2015) Explaining and harnessing adversarial examples. In: International Conference on Learning Representations.
[15] Kurakin, A., Goodfellow, I., Bengio, S. (2017) Adversarial machine learning at scale. In: International Conference on Learning Representations.
[16] Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., McDaniel, P. (2017): Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 [Online]. Available: https://arxiv.org/abs/1705.07204.
[17] Prakash, A., Moran, N., Garber, S., DiLillo, A., Storer, J. (2018) Deflecting adversarial attacks with pixel deflection. arXiv preprint arXiv:1801.08926
[18] Meng, D., Chen, H.: Magnet (2017): a two-pronged defense against adversarial examples. arXiv preprint arXiv:1705.09064