Improving the Transferability of Adversarial Examples with Image Affine Transformation

Heng Yin*, Hengwei Zhang*,†, Zheming Li and Zhilin Liu

PLA SSF Information Engineering University, Zhengzhou 450001, China

*These authors contributed equally to this work
†Corresponding Author: Hengwei Zhang. Email: zhw11qd@aliyun.com

Abstract. Deep learning is widely regarded as a black-box technology. We all know its performance is very good, but we have limited understanding of why it is so good. At present, there are many researches on the interpretability of deep neural network. By studying adversarial example, we can understand the internal semantics of neural network and find the decision boundary with problems, which in turn helps to improve the robustness and performance of neural network and its interpretability. With the development of adversarial example research, more and more adversarial example generation methods are proposed. Although the attack from adversarial example poses a great security threat to the deep learning system, it can also be used as an effective tool to measure the robustness and reliability of the model, and the attack and defense are two mutually promoting processes. Therefore, how to generate adversarial example with stronger attack ability is worth further study. And this study proposes a method named Affine-Invariant Method, aimed to improve the transferability of adversarial examples in black-box environment.

1. Introduction

In recent years, with the rise of deep learning, artificial intelligence has ushered in a new round of development [1]. With the support of big data and powerful computing power, deep neural network plays an increasingly important role in many fields, such as computer vision, natural language processing and speech recognition [2-4]. In some specific application scenarios, its performance has even surpassed the human level, such as image classification [5-6]. With the large-scale application of artificial intelligence technology, its security problems are becoming more and more prominent. For example, in automatic driving, due to the system's wrong judgment of road conditions, it may lead to serious safety accidents. In 2018, the China Academy of Information and Communications issued the white paper on artificial intelligence security, which pointed out the security risks caused by the immature artificial intelligence technology at the present stage, including the technical limitations such as the incomprehensibility of algorithms and strong data dependence, which may bring security risks to cyberspace and national society. It can be seen that improving the interpretability and reliability of artificial intelligence algorithm is an important way to avoid security risks. The concept of model is introduced into artificial intelligence. This model often relies on massive data and powerful computer computing power, trains the model in advance, and then applies it to the actual scene. Facts have proved that the trained artificial intelligence model can achieve good results, but the interpretability and logicality are poor. It is often difficult to find out the internal logic of the model to make decisions artificially. Therefore, the artificial intelligence model may have potential safety hazards that are difficult to be detected by people. With the continuous development of deep neural network in recent
years, artificial intelligence technology has made a major breakthrough in computer vision, especially in the field of image. With the support of big data, neural network shows excellent performance in application scenarios such as target detection, image segmentation and recognition, and has been widely used in practice, such as automatic driving, face payment, etc. Therefore, the security of neural network model has gradually become the focus of research in artificial intelligence security [7]. As the most basic task in computer vision, image classification based on convolutional neural network is vulnerable to the attack of adversarial example, resulting in the failure of deep learning model. In 2013, Szegedy et al. [8] first proposed the concept of adversarial example, that is, by adding the noise that is not visible to the naked eye that is carefully made by human to the image, the neural network can make mistakes in classification with high probability, and the image that is misclassified is adversarial example.

2. Background

2.1. Adversarial Attacks
Counter attack refers to the use of generated adversarial example to attack the neural network, making the model classification error. Counter attack can be classified according to three dimensions. According to the attacker's understanding of the model, it can be divided into white box attack and black box attack. White box attack means that the attacker has all the information including the model structure and internal weight parameters, while black box attack means that the attacker does not understand the model structure and parameters, and can only test the model input and then obtain the model output. Generally speaking, black box attacks can be divided into the following three categories: 1) decision based black box attacks. In this case, although the attacker cannot obtain the internal structure and parameters of the model, he can obtain the model output by inputting the model, that is, querying the model and generating adversarial example based on the corresponding input-output relationship [9-11]; 2) attacks based on alternative models In the method, the attacker cannot obtain the detailed knowledge about the model, but can input the image to the model to obtain the output label, and train an alternative model to simulate the target model to be attacked, and generate adversarial example based on the model [12]; 3) migration attack, which is usually implemented by improving the migration of adversarial example in the white box attack. The first two methods often need a lot of queries on the model, but this is unrealistic in some cases. For example, the online platform usually limits the frequency and number of queries [13-14].

2.2. Adversarial Defenses
Facing the attack of adversarial example, many methods can be used to improve the robustness of the model [15-18], among which the most direct defense method is to carry out adversarial example training, that is, adding adversarial example to the neural network training set to train the model, which is equivalent to making it "recognize" the characteristics of counter sample in advance in the model training stage, so as to improve the robustness of the model. In order to ensure the effectiveness of the method, it is necessary to use adversarial example images with sufficient strength, and the model is required to have strong expression ability. However, some studies have pointed out that no matter how many adversarial examples are added in the training set, there are new adversarial example that can attack the network. Tramèr et al. [19] Proposed the integrated adversarial example training to further improve the robustness of the defense model. Some literatures also suggest that confrontation training can reduce the over fitting of the model and enhance its generalization ability. The methods of modifying adversarial example include random rescaling, random filling, filtering, image enhancement, JPEG image compression and so on. In the aspect of modifying the network structure, Papernot et al. [12] Proposed a defensive distillation method, which can resist the adversarial example samples with small disturbance amplitude. Jin et al. [13] Proposed a kind of feedforward convolution network to prevent adversarial example by adding noise. In the aspect of using external models, Lee et
al. [14] Trained a model to defend against gradient based adversary attacks by using the generated adversary network.

3. Methodology

3.1. Affine Transformation
Affine transformation changes of image include Scale, transform, rotate, reflection, and shear mapping, which feel like the reflection of a graph. The original line is still a straight line after affine transformation. The original parallel lines are still parallel lines after the affine transformation, so that's affine. Affine transformation in the collection of some properties remain the same: (1) convexity; (2) linear: if several points before the transformation in a line, after the affine transformation is still in line; (3) a parallelism: if two lines parallel before transformation, the transformation after the parallel; (4) still collinear ratio invariance: transform the proportion of two previous line, after the transformation ratio unchanged.

Affine transformation is an important transformation in the two-dimensional plane, which has a wide range of applications in the field of image graphics. In the two-dimensional image transformation, it is generally expressed as:

\[
\begin{bmatrix}
  u \\
v \\
1
\end{bmatrix} =
\begin{bmatrix}
a_1 & b_1 & c_1 \\
a_2 & b_2 & c_2 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

3.2. Affine-Invariant Method

Figure 1. Schematic diagram of adversarial example generation based on image affine transformation

Combined with the image affine, this paper proposes an Affine-invariance attack Method (AIM) to optimize the adversarial example, as is shown in figure 1. Similar to other adversarial example generation methods based on data enhancement, AIM can be combined with any iterative gradient-based method, which can be expressed as follows:
\[
\text{arg max}_x J(\theta, A(x), y), \quad \text{s.t.} \| x^{\text{adv}} - x \|_\infty \leq \varepsilon,
\]

where the \( x \) denotes the input image, \( J \) denotes the loss function of the model, \( \theta \) denotes the parameters of the neural network, \( y \) denotes the ground-truth label of \( x \), \( A(\cdot) \) denotes the affine transformation and \( \varepsilon \) denotes the \( \ell_\infty \) norm limitation of the adversarial noise.

4. Experiments
In this section, we have carried out sufficient experiments to verify the effectiveness of the above method. This paper first introduces the experimental data, model and parameter settings, and then shows the results of white-box attack and black-box attack.

4.1. Experimental setup
If normal images cannot be correctly classified by the network, then there is no point in generating an adversarial sample based on these images. Therefore, this paper randomly selects 1000 images from 1000 categories of the ImageNet verification set, each of which belongs to a different category, and all images can be correctly classified by the test network. In the experiment, seven models are studied, four of which are normally trained networks, namely Inception-v3 (Inc-v3), Inception-v4 (Inc-v4), Inception-ResNet-v2 (IncRes-v2) and Resnet-v2-101 (Res-101), the other three are confrontation training networks, namely ens3-adv-incceptionv3 (Inc-v3ens3), ens4-adv-Inception-v3 (Inc-v3ens4) and ens-adv-Inception_resnet-v2 (IncRes-v2ens).

4.2. White-box attack and black-box attack of AIM
We perform adversarial attacks on a single network with I-FGSM, MI-FGSM, DIM and AIM to generate adversarial examples only on the normally trained networks and tested them on all seven networks. The results are shown in Table 1, where the success rate is the model classification error rate with adversarial examples as input.

| Model     | Attack      | Inc-v3 | Inc-v4 | IncRes-v2 | Inc-101 | Inc-v3ens3 | Inc-v3ens4 | IncRes-v2ens |
|-----------|-------------|--------|--------|-----------|---------|------------|------------|--------------|
| Inc-v3    | I-FGSM      | 100.0  | 27.5   | 23.0      | 20.9    | 6.2        | 4.7        | 1.8          |
|           | MI-FGSM     | 100.0  | 54.3   | 50.6      | 44.0    | 13.9       | 13.3       | 6.5          |
|           | DIM         | 99.3   | 70.8   | 66.6      | 60.1    | 15.3       | 16.0       | 7.9          |
|           | AIM(Ours)   | 97.8   | 79.7   | 74.6      | 68.1    | 22.5       | 22.4       | 10.1         |
| Inc-v4    | I-FGSM      | 43.8   | 100.0  | 27.4      | 23.4    | 6.1        | 6.3        | 2.4          |
|           | MI-FGSM     | 69.9   | 100.0  | 57.9      | 54.1    | 19.6       | 17.7       | 8.7          |
|           | DIM         | 76.3   | 99.6   | 69.2      | 65.2    | 23.2       | 19.6       | 11.5         |
|           | AIM(Ours)   | 83.9   | 97.5   | 77.0      | 70.0    | 27.2       | 24.6       | 15.1         |
| IncRes-v2 | I-FGSM      | 46.1   | 35.0   | 99.4      | 30.4    | 7.3        | 6.7        | 4.4          |
|           | MI-FGSM     | 73.5   | 69.3   | 99.5      | 60.0    | 27.0       | 23.0       | 16.7         |
|           | DIM         | 78.4   | 74.8   | 98.4      | 70.7    | 33.0       | 27.5       | 19.0         |
|           | AIM(Ours)   | 85.4   | 83.8   | 96.3      | 78.0    | 45.2       | 37.5       | 27.0         |
| Res-101   | I-FGSM      | 35.1   | 28.3   | 25.1      | 99.5    | 8.4        | 6.7        | 3.7          |
|           | MI-FGSM     | 60.0   | 55.3   | 50.6      | 99.5    | 22.9       | 19.8       | 11.3         |
|           | DIM         | 71.9   | 69.6   | 66.0      | 99.4    | 33.2       | 27.2       | 14.1         |
|           | AIM(Ours)   | 81.4   | 79.0   | 76.4      | 99.0    | 41.9       | 36.5       | 24.1         |
Figure 2 shows the original images and the corresponding adversarial images crafted by AIM.

![Original vs. Adversarial Images](image)

The figure shows that these adversarial examples have little visual change compared with the original images.

5. Conclusions

In this paper, an adversarial example generation algorithm based on image affine invariance is proposed. This is to improve the mobility of adversarial example in the black box model from another point of view, in addition to improving the mobility of adversarial example from the optimization point of view. Since data enhancement is an effective way to improve the generalization performance of the model, the common image affine transformation is used in the gradient attack, and the image is transformed during the algorithm iteration, so as to increase the diversity of the data and improve the generalization performance of adversarial example.

Acknowledgments

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

References

[1] Russell S J, Norvig P. Artificial Intelligence: A Modern Approach[M]. 2003.
[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.
[3] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
[4] Senior A, Vanhoucke V, Nguyen P, et al. Deep neural networks for acoustic modeling in speech recognition[J]. IEEE Signal processing magazine, 2012.
[5] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[7] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars[J]. arXiv preprint arXiv:1604.07316, 2016.

[8] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[9] Wierstra D, Schaul T, Glasmachers T, et al. Natural evolution strategies[J]. The Journal of Machine Learning Research, 2014, 15(1): 949-980.

[10] Alzantot M, Sharma Y, Chakraborty S, et al. Genattack: Practical black-box attacks with gradient-free optimization[C]//Proceedings of the Genetic and Evolutionary Computation Conference. 2019: 11111119.

[11] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against blackbox machine learning models[J]. arXiv preprint arXiv:1712.04248, 2017.

[12] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017: 506-519.

[13] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9185-9193.

[14] Xie C, Zhang Z, Zhou Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2730-2739.

[15] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.

[16] Buckman J, Roy A, Raffel C, et al. Thermometer encoding: One hot way to resist adversarial examples[C]//International Conference on Learning Representations. 2018.

[17] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations[J]. arXiv preprint arXiv:1711.00117, 2017.

[18] Samangouei P, Kabkab M, Chellappa R. Defense-gan: Protecting classifiers against adversarial attacks using generative models[J]. arXiv preprint arXiv:1805.06605, 2018.

[19] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[J]. arXiv preprint arXiv:1705.07204, 2017.

[20] Jin J, Dundar A, Culurciello E. Robust convolutional neural networks under adversarial noise[J]. arXiv preprint arXiv:1511.06306, 2015.

[21] Lee H, Han S, Lee J. Generative adversarial trainer: Defense to adversarial perturbations with gan[J]. arXiv preprint arXiv:1705.03387, 2017.