Black-box adversarial attacks based on random coordinate selection

Kaikun Dong¹, Jiao Shi¹ and Li Guo¹

¹School of Computer Science and Technology, Harbin Institute of Technology at Weihai, Weihai, China

Corresponding author and e-mail: Kaikun Dong, dongkaikun@hit.edu.cn

Abstract. Recent studies show that deep neural networks are vulnerable to adversarial attacks, and the research of adversarial attack has become a hotspot in the field of artificial intelligence security. The decision-based black-box attack is one of the most challenging problems in the field of adversarial attack. Decision-based black-box attacks usually have the problems of high query times and low attack success rate. Several algorithms have been proposed to solve this problem and one of the algorithms presented by Cheng et al. models this attack as an optimization problem. However, this algorithm is based on the method of adding perturbation to the whole pixel, the number of queries is still very high and the convergence perturbation is not necessarily the minimum. In this paper, we present an improved strategy based on stochastic coordinate selection, hoping to improve the query efficiency while reducing the distortion. The convergence of the improved algorithm is analysed, and experiments are carried out on CIFAR-10 and ImageNet datasets. It has been shown that, compared with the original algorithm, our algorithm can find lower distortion using the same number of queries, and the success rate can be improved to different degrees under different query restrictions.

1. Introduction

In recent years, deep learning technology has made great progress driven by massive amounts of data and powerful computing capabilities, especially in the fields of speech recognition, computer vision, and natural language processing. Various products and services based on deep learning are gradually being applied in the industry, such as machine learning systems for self-driving cars and face recognition systems, etc. However, studies have shown[1] that the deep neural network often gives wrong classification results when facing the adversarial samples, that is, it is easy to be attacked by adversarial samples. The adversarial sample is a synthetic image sample constructed by adding a specific distortion to the original image. The human eye can correctly identify its category, but the deep neural network cannot give the correct classification result. The process of constructing the adversarial sample is to adversarial attack.

Adversarial attack is mainly divided into white-box attack and black-box attack. White-box attack can obtain all the information and parameters inside the machine learning model during the attack, and generally generate adversarial samples based on the gradient of a given model. For black-box attack, the structure of the neural network is black box, and the adversarial samples are generated only through the input and output of the model. The white-box attack has high quality and high success rate, but for black-box attack, success rate is low and attack is difficult and challenging. In most cases, it is
difficult for attackers to obtain the specific information of the model, such as the attacker cannot obtain the internal parameter information of the vehicle identification system from the outside of the vehicle. So the white-box attack is not practical, and the black-box attack has more research and application significance.

Researchers have come up with a number of attack approaches. According to different partition rules, attack approaches can be divided into different categories. In terms of attack mode, adversarial attack is divided into white-box attack and black-box attack. From the point of view of target classification, adversarial attack can be divided into targeted attack and untargeted attack. The difference is whether to limit which category the adversarial sample belongs to. From the aspect of learning mode, it can be divided into single-step attack and iterative attack.

In this paper, we consider the most challenging and practical decision-based black-box attack. Simply, the principle of the attack is to generate adversarial samples by adding reasonable noise to the original image. But the perturbation added to the original image may be sparse [2], that is, the perturbation to a portion of the pixel is sufficient to find the adversarial sample. [3] belongs to full pixel distortion and has high complexity. We optimize [3] by means of stochastic coordinate selection, so as to accelerate the efficiency of black-box attack. The feasibility and efficiency of the improved algorithm are verified on two different datasets. Our algorithm can find smaller perturbations and achieve a higher success rate under the restriction of the same number of queries, and we find that the attack time of our algorithm is shorter.

2. Related work
According to the attack principle, black-box attack can be divided into gradient-based attack, score-based attack, transfer-based attack and decision-based attack.

Gradient-based attacks rely on detailed model information or confidence scores, including the gradient information obtained by calculating the loss function. Examples include Deep Fool [4] and the Carlini&Wagner attack [5] etc. Score-based attacks, some attacks that rely solely on the prediction Score of the model (e.g., class probability), include JSMA [6] and the variant of Carlini&Wagner attacks [7]. Transfer-based attacks require information about training data used to train a completely observable alternative model from which adversarial perturbations are synthesized. Decision-based attacks, only the final decision, i.e. the top-1 predicted class is observed. As a result, the attacker can only make queries to acquire the corresponding hard-label decision instead of the probability outputs.

One idea of the algorithm is to test the gradient direction of each point on the sample data vector, and then estimate the iteration step size to calculate and generate adversarial samples. The other is to first convert the attacked black-box model into white-box model by stealing model parameters or model migration, and then use the original white-box attack method to generate adversarial samples.

Several approaches have been proposed to perform decision-based black-box attacks. A commonly used method is the boundary attack proposed in 2017[8], which generates an adversarial sample through a random orthogonal search on the decision boundary. Ilyaset al. [9] proposed the Query-limited attack in 2018, and tried to use output probability scores to transform the hard-label attack into soft-label attack problem, and further into a white-box attack problem. In a paper published in 2019 by Cheng et al. [3], the minimum perturbation problem was transformed into an optimization problem based on the idea of boundary attack.

3. Problem formulation
The black-box attack can be transformed into a minimum optimization problem to find the adversarial sample closest to the original sample. Classification model \( f : \mathbb{R}^d \rightarrow \{1,\ldots,K\} \), \( f(x) \) denotes the output of the model, \( x \in \mathbb{R}^d \) denotes the input of the model, the original sample \( x_0, y_0 \) denotes the original label. The goal of the attack is to find an adversarial example \( x^* \) to satisfy \( f(x^*) \neq y_0 \). We generate adversarial examples by adding perturbation to the original example, while limiting the
perturbation, the purpose is to make the original image and the perturbed image as similar as possible. We can measure the similarity to the adversarial sample by p-norm. The p-norm calculation equation is shown in equation (1). Common norms include \( L_0 \) norm, \( L_2 \) norm, \( L_{\infty} \) norm, etc. Our method uses \( L_2 \) norm measure similarity.

\[
\| x \|_p = \left( \sum_{i=1}^{N} | x_i |^p \right)^{1/p}
\]

Untargeted black-box attack problem is modeled as an optimization problem as in equation (2), and targeted black-box attack problem is modeled as an optimization problem as in equation (3).

\[
\min \| x^* - x_0 \|_p \quad s.t. f(x^*) \neq y_0
\]

\[
\min \| x^* - x_0 \|_p \quad s.t. f(x^*) = t
\]

4. Proposed method
This section explains the algorithm detail of Cheng et al. (2019). Cheng et al. (2019) regard the hard-label attack as a problem of finding the direction of the shortest distance to the decision boundary. The distance is evaluated by binary search and model access results, and the search direction is updated by the gradient calculation equation.

Specifically, the objective function is the distance from the original image to the decision boundary along a specific direction. The objective function for untargeted attacks is shown in equation (4), and for targeted attacks is shown in equation (5). Symbol \( \theta \) denotes searching direction, \( \lambda \) denotes perturbation and \( g(\theta) \) is the distance from the original image \( x_0 \) to the nearest adversarial sample in the direction \( \theta \). The optimization problem expression is shown in equation (6).

\[
g(\theta) = \arg \min_{\lambda>0} (f(x_0 + \lambda \frac{\theta}{\|\theta\|}) \neq y_0)
\]

\[
g(\theta) = \arg \min_{\lambda>0} (f(x_0 + \lambda \frac{\theta}{\|\theta\|}) = t)
\]

\[
\min_{\theta} g(\theta)
\]

The adversarial example \( x^* \) are generated by equation (7), where \( \theta^* \) is the optimal solution of equation (6). The estimated gradient equation is shown in equation (8), where \( H \) is a random Gaussian vector, and \( \beta > 0 \) is a small smoothing parameter.

\[
x^* = x_0 + g(\theta^*) \frac{\theta^*}{\|\theta^*\|}
\]

\[
\hat{g}(\theta) = \frac{g(\theta + \beta H) - g(\theta)}{\beta}, \mu
\]

5. Approach
When updating the direction of the disturbance, Cheng et al. added Gaussian noise of the same dimension as the original image. The number of pixels modified is quite large, so this modification may be excessive resulting in failure to find smaller distortions. The difference in our approach is that distortions are performed on a small number of pixels to construct adversarial samples.
5.1. Stochastic coordinate selection
Each pixel in the image is regarded as a dimension, so that the dimension of the whole search space will be very high when the whole pixel is perturbed, which often means that the search efficiency is very low. For this reason, we use an acceleration strategy called stochastic coordinate selection. By randomly selecting some coordinate points, the noise of the selected coordinate position is retained, and the noise depth of the Gaussian noise of the unselected position is set to 0. In other words, each sample is only taken in some dimensions (that is, in some pixels), not in all dimensions. The specific algorithm description is shown in Algorithm 1.

Algorithm 1: The Stochastic coordinate selection algorithm

Input The dimensions of the original image $n \in \mathbb{N}_+$, the number of coordinates $m \in \mathbb{N}_+$ for stochastic coordinate selection;
Randomly sample $u$ from a zero-mean White Gaussian Noise: $u \in \mathbb{N}(0, 1)$
Select $m$ coordinates from $n$ with the same probability;
Set the non-selected coordinates of $u$ to 0;
Return $u^*$

5.2. Black-box attack based on stochastic coordinate selection
When sampling Gaussian vector, we adopt the random coordinate selection method in section 5.1, and use the same gradient calculation equation as Opt to calculate the gradient. The improved algorithm is described as Opt-Low-Attack (Optimization-based Low-dimension Attack) in Algorithm 2.

Algorithm 2: Opt-Low-attack

Input attack model $f$, original image $x_0$, initial $\theta_0$;
For $t = 1, 2, \ldots, T$ do
Randomly choose $u$ using Algorithm 1;
Evaluate $\nabla g$ using the same algorithm in Cheng et al.(2019);
Update $\theta_{t+1} = \theta_t - \eta_t \nabla g$;
Return $x_0 + g(\theta_f)\theta_f$

6. Experiments

6.1. Experimental settings
In our experiments, we use the same datasets as Opt algorithm, namely CIFAR-10 and ImageNet. Among them, CIFAR-10 uses CNN with four convolutional layers, two maximum pooled layers, and two fully connected layers. For the ImageNet, torchvision provided the pre-training network RESNET-50. All models were trained using PyTorch.

This paper only carries out the untargeted attack under the $L_2$-norm perturbation, but it can also realize the attack under other norm perturbation or targeted attack. It only takes a small amount of tweaking when attacking. In each attack, 100 images are randomly selected for CIFAR-10, while 25 images are randomly selected for ImageNet. Because for high dimensional dataset like ImageNet, the attack time is proportional to the data dimension, and the attack is very time-consuming. The original images were all from the validation set part of the dataset, and we attack only those images that are correctly classified by the model and skip those that were misclassified. During the experiment, we ensured that the comparison with the original algorithm was carried out under the same experimental setting through fixed random seeds.

About hyperparameters, our method has only one hyperparameter which is the number of random coordinate points. We randomly take several groups of data and carry out a large number of
experiments. For CIAFR-10, the number of random coordinate points was selected as 2000 and 2500, and the improved algorithm was Opt-Low1-attack and Opt-Low2-attack respectively. For ImageNet, we selected 80000 and 100000 coordinate points, and the improved algorithm was Opt-Low3-attack and Opt-Low4-attack respectively.

In this paper, the size of distortion, query times and success rate of attack are used to evaluate the performance of the algorithm. As for the size of distortion, we calculate the median of the distortion. Compared with the minimum distortion and the average distortion, the median distortion is a suitable index to evaluate the performance of the algorithm. Number of queries is the number of accesses to the model. For success rate of attack, under a certain limit of number of queries, the distortion added to the original sample is lower than the given perturbation threshold, the attack is successful.

6.2. Experimental results

We reproduced the algorithm in [3] and did comparative experiments on CIFAR-10 and ImageNet. For CIFAR-10, we randomly selected 5 images with different labels from the experimental results and drew an image where the distortion size changed with number of queries, as shown in Figure 1.

We find that for CIFAR-10, with smaller query number limit, our method generates lower distortion when query number is the same, and in some cases, our method can converge to a smaller distortion.
Table 1. Untargeted attack-Comparison of median $L_2$ distortion and SR(success rate) obtained using a given number of queries and $\varepsilon = 0.5$ for different attacks in CIFAR-10.

| Queries | Median $L_2$ | SR  |
|---------|--------------|-----|
| Opt-attack | 4000 | 0.7735 | 37.0% |
|          | 8000 | 0.4353 | 53.0% |
|          | 14000 | 0.3312 | 61.0% |
|          | 4000 | 0.7672 | 41.0% |
| Opt-Low1-attack | 8000 | 0.4291 | 58.0% |
|          | 14000 | 0.2892 | 68.0% |
|          | 4000 | 0.7574 | 43.0% |
| Opt-Low2-attack | 8000 | 0.4089 | 56.0% |
|          | 14000 | 0.2845 | 70.0% |

As can be seen from Table 2, under the same query number limit, our method generates lower distortion and increases success rate to different degrees. At the query number limit of 14000, our attack success rate reaches 70%. Particularly for ImageNet dataset, Opt-Low4-attack reaches distortion below 5.0 in about 30k queries while Opt-attack need more than 150k queries for the same.

Table 2. Untargeted attack- Comparison of median $L_2$ distortion and SR(success rate) obtained using a given number of queries and $\varepsilon = 3$ for different attacks in ImageNet.

| Queries | Median $L_2$ | SR  |
|---------|--------------|-----|
| Opt-attack | 4000 | 63.55 | 0% |
|          | 30000 | 10.28 | 14.0% |
|          | 150000 | 4.08 | 20.0% |
|          | 4000 | 61.25 | 0% |
| Opt-Low3-attack | 30000 | 6.12 | 24.0% |
|          | 150000 | 4.78 | 30.0% |
|          | 4000 | 58.83 | 0% |
| Opt-Low4-attack | 30000 | 4.36 | 24.0% |
|          | 150000 | 4.37 | 40.0% |

7. Conclusions
We proposed stochastic coordinate selection to improve Opt algorithm and realize a more efficient black-box attack algorithm. Through experiments, we prove the effectiveness and convergence of the algorithm. Under the same limited conditions, the distortion generated by our method is smaller and the success rate of attack is improved to different degrees. In addition, the improved strategy can provide a new idea for black-box optimization, that is, there are some limitations when generating noise, and low-dimensional noise can be generated according to a certain distribution or probability, so as to carry out local pixel perturbation.

References
[1] Aishan L, Xianglong L and Jiaxin F 2019 Perceptual-Sensitive GAN for Generating Adversarial Patches Proceedings of the AAAI Conference on Artificial Intelligence pp 28–35
[2] Su J, Vargas D V and Sakurai K 2019 One pixel attack for fooling deep neural networks IEEE Transactions on Evolutionary Computation pp 828-841
[3] Minhao C, Thong L, Pin-Yu C, Huan Z, JinFeng Y and Cho-Jui H 2019 Query-efficient hard-label black-box attack: An optimization-based approach International Conference on Learning Representations
[4] Seyed-Mohsen Moosavi-Dezfooli, Fawzi A and Frossard P 2016 Deepfool: a simple and accurate method to fool deep neural networks IEEE conference on computer vision and
pattern recognition pp 2574-2582

[5] Carlini N and Wagner D 2016 Towards Evaluating the Robustness of Neural Networks IEEE Symposium on Security and Privacy pp 39-57

[6] Narodytska N and Kasiviswanathan S 2016 Simple black-box adversarial perturbations for deep networks. arXiv preprint arXiv:1612.06299

[7] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh 2017 Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. arXiv preprint arXiv: 1708.03999

[8] Brendel W, Rauber J and Bethge M 2017 Decision-based adversarial attacks: Reliable attacks against black-box machine learning models arXiv:1712.04248

[9] Ilyas A, Engstrom L, Athalye A, et al 2018 Black-box Adversarial Attacks with Limited Queries and Information Black-box adversarial attacks with limited queries and information International Conference on Machine Learning pp 2142–2151