Research on Generating Adversarial Examples in Applications

YangZhao, Juan Wang*, Yingjiang Liu
School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China

*wangjuan@cuit.edu.cn

Abstract. The deep neural network is a highly expressive model, which plays an extremely critical role in modern artificial intelligence applications. Adversarial samples can change the prediction results of neural networks by imposing imperceptible perturbations on the original images, which brings new challenges to deep learning. In this paper we summarize the methods of generating adversarial samples in recent years, intuitively feel the development of adversarial samples from the time of publication, and briefly classify them from the perspective of algorithm principles. At the same time, for practical applications, the advantages and disadvantages of the algorithm are analyzed and summarized, and the conditions for the algorithm to be suitable for application are proposed: no need to know the specific structure of the algorithm, high quality of the adversarial samples and convenient migration. The analysis points out that among the typical methods, MI-FGSM, ONE-PIXEL, and methods using GAN are more practical and worthy of further study.

Keywords: Deep Neural Network; Adversarial Examples; Perturbation

1. Introduction

In the era of big data and artificial intelligence, the wide application of deep learning technology in various fields has brought convenience to people's lives and also caused many security issues. As early as 2013, Szegedy et al. first proposed the concept of "adversarial samples": by optimizing the input to maximize the prediction error to find a non-random perturbation, imposing this imperceptible perturbation on the test image can arbitrarily change the network prediction. The generation of adversarial samples has great practical significance for both offensive and defensive aspects of deep learning.

The existing researches on adversarial sample generation algorithms mostly focus on the theoretical research field, and few researchers pay attention to its application. In the application in the publicly published literature, Thys[1] et al. put people on clothes with anti-patch to make the pedestrian detection classifier misclassify;Evtimov[2] et al. used various methods to construct adversarial examples of stop signs and misclassified the road sign recognition model. However, the generation of these adversarial samples either needs to know the internal parameters of the algorithm, or is limited to a special field, the sample generation is not easy, and the generation method has strong limitations.

This paper is oriented to real-world applications and summarizes the advantages and disadvantages of typical generated adversarial samples from the perspective of the application. And put forward the
standard to measure the application prospect of the algorithm, and then analyze which algorithm is the best application prospect in the existing confrontation sample generation method, which provides a good reference for the application of the confrontation sample generation algorithm.

2. Adversarial example development
This paper explains the development of adversarial samples from the perspective of the publication time of various algorithms in order to get a more intuitive experience of the development of adversarial samples[3]. The development of adversarial examples in terms of time is shown in the following figure:

![Fig.1 The development of adversarial examples in terms of time.](image)

Szegedy et al. found that by maximizing the prediction error of the network to discover some imperceptible perturbation, the network can misclassify images. This is the first time that the concept of adversarial examples has been proposed, and a formulaic description of the way to generate efficient adversarial examples: minimize $\|p\|_2$ subject to $(p$ is the perturbation of the data $): 1. f(x + p) = l; 2. x + p \in [0,1]^m$. And an algorithm for generating adversarial examples: L-BFGS has also been proposed.

Goodfellow et al. proved that the non-linear characteristics of neural networks led to the generation of adversarial examples, and proposed a method to quickly generate adversarial examples, FGSM. This method is almost all the basic methods based on gradient attacks in the later stage.

Papernot et al. proposed a method of adversarial sample generation based on neural network types called JSMA. This method requires the calculation of the number of forward guides and is a more powerful method than the method based on gradient descent. The perturbation norm of JSMA is limited to L0, in order to minimize the modification of the original sample pixels. There is a flaw in
the experiment. On a data set with too many pixel values, it is computationally difficult to search for two locations, and the performance of searching only one location is too poor, so this algorithm is not recommended. For example, the ImageNet dataset.

Moosavi-Dezfooli et al. proposed the DeepFool algorithm for effective calculation of counter perturbation. Compared with FGSM, this method can generate smaller disturbance values through iterative calculation.

Kurakin et al. proposed the basic iterative method I-FGSM. The basic idea is to optimize the large-step operation to increase the loss function of the classifier through multiple small steps to increase the loss function, thereby performing image perturbation. In order to get a more refined confrontation sample.

Carlini and Wagner proposed a new attack method, C&W, and successfully attacked the distillation network, proving that it did not significantly improve the robustness of the neural network.

Moosavi-Dezfooli et al. proved the existence of universal perturbation and proposed an algorithm for finding universal perturbations, Universal adversarial perturbations.

Baluja and Fischer proposed a new method to generate adversarial samples, which is different from the previous method to solve an optimization problem. By effectively training a feedforward network in a self-supervised manner to generate adversarial samples, it can be used to attack one or more target networks, And call this kind of network ATNs.

Sarkar et al. proposed two black-box attack methods, UPSET and ANGRI, for targeted attacks.

Cisse et al. proposed a new and flexible method, Houdini, to generate adversarial examples. Unlike other methods, it has its own loss function.

Karmon et al. proposed the LaVAN method, which adds perturbation to the local position of the image when the noise is visible, so as to generate good confrontation samples.

Aiming at the problem of the low success rate of existing black-box confrontation attacks, Dong et al. proposed an iterative algorithm based on momentum to enhance confrontation attacks, and experiments have proved that the effect is better. By integrating the momentum term into the iterative process of the attack, this method can stably update the method, and get rid of the bad local maximum during the iterative process, thereby generating more transitive adversarial samples.

Brown et al. proposed a method to create a universal, robust, and targeted "patch" against images in the real world. Compared with the traditional modification of the target image for confrontation, this method initiates an attack by completely replacing part of the image with a practical confrontation image patch.

Su et al. analyzed in a very limited situation: that is, only one pixel can be modified and proposed a single-pixel confrontational perturbation generation method based on differential evolution. Compared with other methods, by applying differential evolution to generate adversarial sample images, the probability of finding the global optimal solution is higher, and only a small amount of information is needed to carry out a black box attack.

Goodfellow[4] et al. in October 2014 proposed a framework GAN(Generative Adversarial Networks) for estimating the generative model through the confrontation process.Xiao[5] et al. proposed AdvGAN, a method that uses generative adversarial networks to generate adversarial samples. The core is to map the clean sample into a confrontational perturbation through the GAN generator, and then add it to the clean sample. The discriminator is responsible for identifying the input sample as a confrontational sample. It has better results than traditional methods in white box testing and black-box testing. Based on AdvGAN, Mangla[6] proposed a method with a higher attack rate, AdvGAN++, whose core is to introduce the hidden layer vector in the classifier as the input of GAN to generate adversarial samples.

3. Application senario analysis

From the "adversarial sample" was proposed to the present, a large number of algorithms for generating confrontation have been proposed, including all the above algorithms. There have been some reviews of adversarial examples. They mainly focus on summarizing and analyzing the
algorithm from the attributes of various algorithms, such as whether the algorithm is mainly used for white-box attacks or black-box attacks, targeted attacks, or non-targeted attacks, general or specific, perturbation norm limits, etc. Summary of attributes of various attack methods listed is shown in the following table:

| Methods                              | Black(B)/White(W) | Targeted(T)/Non-targeted(N) | Universal(U)/Specific(S) | Perturbation norm |
|--------------------------------------|-------------------|-----------------------------|--------------------------|-------------------|
| L-BFGS                               | W                 | T                           | S                        | \(l_\infty\)     |
| FGSM                                 | W                 | T                           | S                        | \(l_\infty\)     |
| JSMA                                 | W                 | T                           | S                        | \(l_0\)           |
| DeepFool                             | W                 | N                           | S                        | \(l_2/l_\infty\) |
| I-FGSM, ILCM                         | W                 | N                           | S                        | \(l_\infty\)     |
| C&W                                  | W                 | T                           | S                        | \(l_0l_2\)       |
| Universal adversarial perturbations  | W                 | N                           | U                        | \(l_2/l_\infty\) |
| ATNs                                 | W                 | T                           | S                        | \(l_\infty\)     |
| UPSET                                | B                 | T                           | U                        | \(l_\infty\)     |
| ANGRI                                | B                 | T                           | S                        | \(l_\infty\)     |
| Houdini                              | B                 | T                           | S                        | \(l_2l_\infty\) |
| LaVAN                                | W                 | N                           | U                        | \(l_\infty\)     |
| MI-FGSM                              | B                 | N                           | S                        | \(l_2l_\infty\) |
| Adversarial Patch                    | B                 | N                           | U                        | \(l_\infty\)     |
| ONE-PIXEL                            | B                 | N                           | S                        | \(l_0\)           |
| AdvGAN                               | B                 | T                           | S                        | \(l_2\)           |
| AdvGAN++                             | B                 | T                           | S                        | \(l_2\)           |
| Natural GAN                          | B                 | T                           | S                        | \(l_2\)           |
| AdvFaces                             | B                 | N                           | U                        | \(l_2\)           |

**Tab.1 Summary of attributes of various attack methods listed**

OpenAI believes that there are stable adversarial samples in the physical world. Kurakin et.al have discussed and proved that adversarial samples are feasible in the real world. The algorithm will eventually be more valuable when it is applied in practical applications. The previous article focused on the theoretical comparison and lacked analysis of the application scenarios and advantages and disadvantages of the algorithm. This paper is oriented to the practical application of the applicable fields and application scenarios of the analysis algorithm.

First of all, we believe that the black box generation method is more practical because it does not need to know the internal structure of the algorithm. The black box method means that the attacker does not need to obtain internal information such as the target model structure or parameters to attack the model. It can be found that from the earliest L-BFGS to the most recent ONE-PIXEL, the adversarial generation algorithm has gradually transformed from the application of white-box attacks to the application of black-box attacks. Although white-box attacks and black box attacks are equally popular in research, starting from practical applications: almost all neural network applications that have landed no longer disclose their internal structure parameters, so black-box algorithms are more applicable and have high application prospects. Based on the white box algorithm, black box algorithms include UPSET&ANGRI, Houdini, MI-FGSM, Adversarial Patch and ONE-PIXEL, etc.

Secondly, the quality of the generated adversarial samples greatly affects its application prospects and applicable fields. According to the research of the iterative algorithm, the quality of the adversarial samples is high. The quality of the adversarial sample or the strength of adversarial is reflected in the deception of the human eye and the deception of the classifier. The deception of the
human eye is invisible to the human eye, which is mainly produced by tiny pixel perturbation. For example, full pixel perturbation, partial pixel perturbation. The methods represented by LaVAN and Adversarial Patch, the changes can be observed by the human eye and are easy to be found, the application prospect is not good and it is only suitable for scenes without human eyes. The deception for the classifier is mainly applied to the deception of various published models, without the participation of people, so the impact on people can be ignored. For classifier spoofing, optimization is mainly carried out by iterative and gradient descent methods, such as MI-FGSM and ONE-PIXEL, which have a wide range of applications.

Finally, the generated confrontation algorithm with good mobility, the samples generated by it can be easily migrated from one application scenario to another. The application prospects of this algorithm are also good and the application range is wide. At present, the adversarial sample generation algorithm using GAN shows good mobility.

In summary, we get this table is shown in the following:

| Applicability Evaluation Standard | Good evaluation | Good overall evaluation |
|----------------------------------|-----------------|-------------------------|
| No need to know the internal structure of the algorithm (black box) | UPSET&ANGRI, Houdini | MI-FGSM, ONE-PIXEL, AdvGAN, AdvGAN++, Natural GAN |
| Quality of adversarial examples | MI-FGSM, ONE-PIXEL, AdvGAN, AdvGAN++, Natural GAN | Methods using GAN (such as AdvGAN) |
| Migration                       | AdvGAN, AdvGAN++, Natural GAN, AdvFaces |

Tab.2 Application analysis of various attack methods listed

4. Conclusion
Previous reviews were mostly limited to analyzing and summarizing the adversarial attack algorithms from the principle in order to produce better adversarial examples. This paper summarizes the advantages and disadvantages of existing typical counter-attack algorithms from the application point of view and proposes the conditions that a suitable algorithm needs to meet: no need to know the specific structure of the algorithm, high-quality countermeasure samples, and convenient migration. Finally, based on the analysis of these conditions, it is concluded that among the existing typical confrontation generation algorithms, MI-FGSM, ONE-PIXEL, and methods using GAN (such as AdvGAN) have good application value, and it is worth further studying their engineering application schemes.

Acknowledgments
The work is supported by Sichuan Science and Technology Program, No.2019YFG0425 and No.2019YFG0186.

References
[1] Thys S, Van Ranst W and Goedemé T 2019 Fooling automated surveillance cameras: adversarial patches to attack person detection 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp49-55
[2] Evtimov I, Eykholt K, Fernandes E, Kohno T, Li B, Prakash A, Rahmati A and Song D 2018 Robust physical-world attacks on deep learning models 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)
[3] Wenwen P, Xinyu W, Mingli S and Chun C 2020 Survey on generating adversarial examples J. Journal of Software 31 pp 67-81

[4] Goodfellow IJ, Pouget-Abadie J, Mirza M, et al 2014 Advances in neural information processing systems Advances in Neural Information Processing Systems 3 pp 2672-2680

[5] Chaowei X, Bo L, Junyan Z, Warren H, Mingyan L and Dawn S 2018 Generating adversarial examples with adversarial networks J. IJCAI

[6] Mangla P, Jandial S, Varshney S and Balasubramanian V 2019 AdvGAN++: Harnessing latent layers for adversary generation J. ICCV