Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/17/20230950

Review on adversarial attack techniques of DNN

Lize Chen

China University of Geosciences, Wuhan, 430074, China

lchen03@outlook.com

Abstract. Deep learning models have excelled in both academia and practice in recent years and have made many developments in various areas. Research has shown that they are inherently vulnerable to attacks by adversarial samples that make them misleading. By studying adversarial attack samples in the field of deep learning security, not only the potential attacks on the models can be reduced, but also their properties can be used to make further improvements to deep learning models. This paper reviews the existing results of adversarial attack techniques for deep neural networks. Firstly, the definition, classification criteria, and development of adversarial attacks are introduced, then the classical white-box and black-box attack methods at the present stage are compared and analyzed, and finally, a summary and outlook are made.

Keywords: deep learning, adverserial attack, sample, migration white-box-attack, black-box-attack.

1. Introduction

In the third decade of the 21st century, deep learning has been broadly applied in various area, including computer vision, natural language processing, speech recognition, etc. Especially in the task of image recognition and classification, deep learning has high accuracy, and in recent periods, deep learning in natural language processing even performs close to human working ability. However, once faced with targeted attacks on deep learning models, those perturbations that are almost imperceptible to humans often make all kinds of deep learning models perform poorly.

In order to build safe and reliable deep learning systems and further optimize the efficiency of DNN models based on the problems generated by the attacks, this paper summarizes and concludes the existing methods of adversarial attacks. This paper introduces the relevant definitions, classification criteria, and development of adversarial attacks and analyzes the methods of adversarial attacks based on the differences in attack methods, thereby giving a clear explanation of the development of deep learning-based adversarial attack techniques in recent years and the future development prospects of adversarial attacks.

2. The basic concept of adversarial attacks

2.1. Proposal of adversarial attacks

Deep learning refers to the method by which a model goes through a data set and learns the intrinsic patterns and representation hierarchy of the sample data. In the exploration of deep learning interpretability by Szegedy et al., it was found that deep learning appears highly vulnerable in samples with specific perturbations added [1]. That is, an attacker can lead a deep learning model to make a wrong classification decision by adding subtle changes to the source data that are difficult for humans to recognize [1]. Such samples are referred to as adversarial samples.

2.2. Relevant concepts in adversarial attacks

2.2.1. Adversarial attack. An attack in which a deep learning model gives false outputs with high confidence by generating adversarial samples to evade deep learning-based detection services is known as an adversarial attack.

2.2.2. Sample distance metric. The sample distance metric, or perturbation, is the distance metric between adversarial samples and original samples. Using the sample distance metric, which is defined as shown in Formula 1, the similarity of the attacking samples can be learned.

$$||\delta_p|| = (\sum_{i=1}^{n} |\delta_i|^p)^{\frac{1}{p}} \qquad (1)$$

where p stands for different paradigm distances, for example, L0 is the number of antagonistic perturbations, L2 is the Euclidean distance between the original and antagonistic samples, and L∞ denotes the maximum change intensity of the antagonistic perturbations, of which the most commonly used is the L2 distance. Generic perturbation: a perturbation that can be superimposed on any image, where generic refers to having image imperceptibility for any image, not to the migratory nature of generic perturbations [2].

2.2.3. Other relevant concepts. Almost undetectable means that the adversarial samples have a small impact on human perception while misleading the classifier of the model, insufficient to change human decision-making.

Transferable means that the adversarial sample also has a certain degree of ability to attack other models other than the one that generated it [3].

Fooling rate means that the proportion of samples in the total sample that causes a change in their classification by the target model after image perturbation.

Dataset means that the data is used in the training of samples to evaluate and compare the performance of attack methods. Common datasets in adversarial attack and defense are ImageNet, MNIST, and CIFAR-10, which are widely used and often used as the base training object for attack methods.

2.3. Classification of adversarial attack

Based on attack effect classification, adversarial attack samples can be divided into directed and undirected adversarial attacks; based on perturbation range classification, they can be divided into global pixel perturbation attacks and partial pixel perturbation attacks; based on attack frequency classification, they can be divided into single attacks and iterative attacks;

based on attack cost classification, they can be divided into white-box attacks and black-box attacks. In this paper, several classical and current cutting-edge counterattack methods based on attack cost classification are introduced.

3. White-box attacks

The white-box attack is an attack that generates an adversarial sample against a network model when the complete model, including the network structure, activation function type, weights and hyperparameters, training data, etc., is known to the researcher [4]. By being familiar with the study model's structure and specific parameters for the various levels, the model's inputs can be controlled or even modified at the bit level by the attacker. The white-box attack is an easy-to-implement sample attack scheme, but the application scenarios are relatively limited due to the difficulty of obtaining the internal knowledge of deep learning models in most scenarios.

3.1. Optimization-based attack methods

3.1.1. L-BFGS. L-BFGS was the first algorithm designed to generate adversarial samples to attack deep learning models. It is used by Christian et al. in the interpretability exploration of deep learning [4]. Its goal is to find an imperceptible or imperceptible minimum input perturbation in the constraint space of the input, and its basic principle is illustrated by Formula 2:

$$\min c||\delta|| + J_{\theta}(x', l') \tag{2}$$

The elements of the input sample x are all regularized to between [0,1], and c is a constant greater than 0. The above optimization process is performed in an iterative form and the parameter c is gradually made larger by linear lookup until the adversarial samples are found.

The quality of the adversarial samples generated by L-BFGS is very dependent on the selection of the appropriate parameter c. Therefore, the method usually requires a lot of time to find the appropriate parameter c, which is one of the most important issues to be considered in optimization-based attacks and limits their performance.

The L-BFGS attack on the then state-of-the-art image classification models AlexNet and QuocNet successfully makes the models misclassify a large number of images. And due to the limitation of L2 parametric distance, the adversarial samples generated by L-BFGS are visually similar to the original unprocessed pre-input images. One of the key points is that L-BFGS converts the difficult optimization problem of generating adversarial samples into a box constraint form for the first time.

3.1.2. C&W. In order to avoid attackers from directly contacting neural network models and performing sample adversarial attacks, Papernot et al. came up with defensive distillation. It is the migration of complex models obtained from the current neural network learning to a neural network with a simpler structure [5].

The C&W attack proposed by Carlini et al. can defeat the model after defensive distillation and misclassify the model with high confidence [6]. The authors also propose effective attacks under different paradigms, and they experimentally verify that the model is best attacked under the L2 paradigm, which is perturbed as shown in the following formulas.

$$\min ||\delta||_2 + c \cdot f(x') \tag{3}$$

$$\text{s.t. } x' = x + \delta \in D \tag{4}$$

$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t) - k \tag{5}$$

$Z(x')_i$ is the output of the neural network softmax. The perturbations added to the adversarial samples produced by the C&W attack perform quite well visually with almost imperceptible characteristics and the sample attack performance of C&W is also excellent, and the C&W attack can still attack successfully and efficiently in the classification model with the addition of defensive distillation. However, since C&W is also an optimization-based attack, it takes a large amount of time to update the parameter c, which is not as fast as other attacks, so it is rarely used to generate counter samples in some AI counter sample competitions with optimization-based approaches such as C&W.

3.2. Gradient-based attack methods

Goodfellow et al. put forward the Fast Gradient Sign Method (FGSM), a method with a core idea to add perturbations along the opposite direction of the negative gradient [7]. In the FGSM, the scale of the perturbation is controlled by:

$$\eta = \epsilon \mathrm{sign}(\nabla_x J(\theta, x, y)) \tag{6}$$

Where η is the model's parameters, θ is the model's input, y is the category that corresponds to the sample x true, and J is the training model's loss function. Goodfellow argues that the linear nature of neural networks in high-dimensional spaces is the real reason for the existence of adversarial samples, rejecting Szegedy's explanation of why neural networks are susceptible to these attacks [4]. Additionally, the concept of adversarial training was offered.

However, this is an undirected attack that cannot perform a directed attack and can only cause the model to be incorrect. Also, this assault is not powerful, and the additional annoyances are effortlessly sifted through in the preprocessing phase of the picture.

BIM (I-FGSM) can be regarded as an improved version of FGSM due to the fact that FGSM performs a one-time change in the size of ϵ for all its pixels. and FGSM adds only single-step perturbation.

Based on this, Kurakin et al. built BIM, which is initialized to random noise, finds perturbations for each pixel point by iteration, and recalculates the gradient direction after each step, which can construct perturbations with less impact relative to the original samples compared to FGSM, but is weak in model migration [8].

The MI-FGSM proposed by Dong et al. adds momentum m to the BIM, similar to the momentum method in optimization methods, to stabilize the updated gradient direction by adding historical inertia to the direction of the parameter update, and thus improve the mobility of the antagonistic samples [9].

The VMI-FGSM proposed by Wang et al. fine-tunes the current gradient by the gradient variance during the previous iterations on the basis of the BIM, with the aim of stabilizing the direction of the gradient update and avoiding the poor results caused by the local optimum; and the VMI-FGSM performs well in sample mobility and can migrate to unknown black boxes for sample attacks [10].

The SMI-FGSM proposed by Wang et al. introduces the gradient in the image's spatial domain and stabilizes its direction by incorporating momentum accumulation from the temporal domain into the spatial domain and taking into account contextual gradient information from various regions [11]. According to the experimental results, the approach achieves the highest migration success rate compared to other advanced methods for several mainstream undefended and defended models.

Table 1. Development of the FGSM method.

Proposed Date	Name	Method Features
2014	FGSM	proposed the reason for the existence of adversarial samples and pioneered gradient-based sample perturbation
2016	BIM(I-FGSM)	Iterative perturbation to attack the model with smaller perturbation
2018	MI-FGSM	Add momentum m to iterative perturbation with some sample migration
2021	VMI-FGSM	Good sample mobility by stabilizing the gradient of variance during iteration
2022	SMI-FGSM	Stabilizes the gradient direction by introducing contextual information and achieves the highest migration success rate so far

AS shown in Table 1, the FGSM has achieved considerable and long-term development and has been widely used in the field.

3.3. Boundary-based attack methods

3.3.1. Deep fool. Deep Fool, proposed by Moosavi-Dezfooli et al. defines the robustness of samples and models for the first time and proposes a simple and effective method to compute the minimum perturbation sufficient to change the labels and to have high attack accuracy by iterative linearization based on the classifier [12].

After iteration, Deep Fool is able to obtain an approximation of the minimum value of the antagonistic perturbation A as shown in the following formulas:

$$\delta_*(x_0) = \underset{\delta}{\operatorname{argmin}} ||\delta||_2 \tag{7}$$

$$\mathrm{s.t.\,sign}(f(x_0 + \delta)) \neq -\frac{f(x_0)}{||\omega||_2^2} \tag{8}$$

where f(x) is the reflection classifier, and f(x)= ω^Tx+b. The authors also analyze and demonstrate the unreasonableness of past algorithms to verify classifier robustness and suggest that using inappropriate algorithms (e.g. FGSM) can over-evaluate classifier robustness. and may lead to incorrect learning results. However, the Deep Fool attack is to make the original samples cross the decision boundary of the classifier at a minimum distance to form an adversarial sample, and thus cannot mislead the deep learning classifier to the specified class, i.e., it does not have the ability of directed attack.

3.3.2. UAP. Building on Deep Fool, Moosavi-Dezfooli et al. found that deep learning models exhibit a general adversarial perturbation, independent of the sample input [13]. This perturbation is related to the structure of the target model and the characteristics of the dataset. When this adversarial perturbation is added to a series of input samples of the data, most of the obtained adversarial samples are able to perform sample attacks.

The UAP attack algorithm generates generic adversarial perturbations on a small number of sampled data points by iterative computation. The UAP attack algorithm uses the Deep Fool algorithm for solving the perturbation computation during the iterative process. Eventually, the data points are pushed to the other side of the model decision boundary after several iterations to counter the attack.

Because UAP namely exists in images and also in various neural networks. So UAP can not only counterattack against unknown samples but also counterattack across models, and it has a relatively high success rate.

3.4. GAN-based attack method

3.4.1. AdvGAN. The AdvGAN proposed by Xiao et al. introduces GAN to the area of sample adversarial attacks for the first time by generating an adversarial perturbation through a generator and adding it to the samples [14]. And the discriminator is mainly responsible for discriminating whether the input is an adversarial sample generated by the generator or a clean sample. The trained AdvGAN network can transform and generate random noise into effective adversarial samples, turning the whole training process into a GAN gaming process.

3.4.2. AdvGAN++. AdvGAN++ proposed by Mangla et al. uses the classifier's hidden layer vector information as input to the GAN in order to produce adversarial samples on the basis of AdvGAN [15]. The attack uses the potential feature map of the original image as the features before adversary generation.

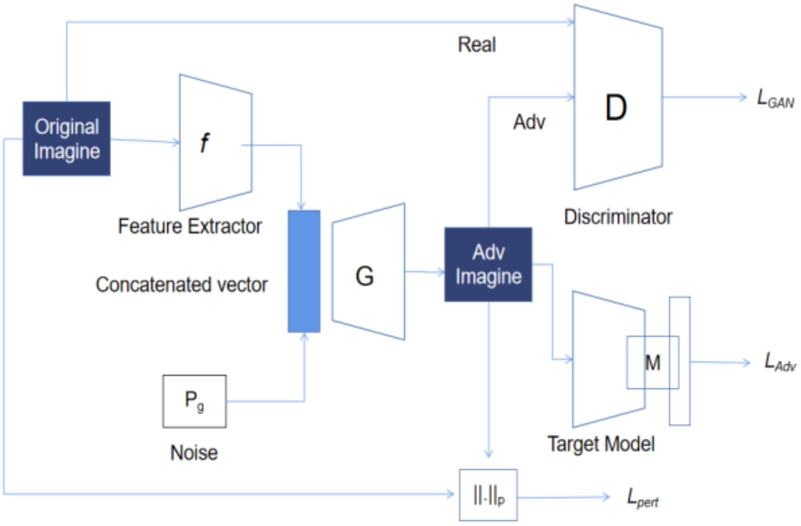

Figure 1. Principle of AdvGAN++.

As Figure 1 shows, by solving the min-max game, AdvGAN++ can obtain the optimal parameters of G and D. Thus, the training process ensures that adversarial samples close to the input distribution are learned to be generated in order to exploit the sensitivity of potential features against perturbations.

The latent features are more vulnerable to adversarial noise than the input image. AdvGAN++ reduces the training time, and meanwhile, it increases the attack success rate. And this attack method has a higher success rate of adversarial sample attack models generated by AdvGAN++ than AdvGAN in all models with defenses.

3.5. Summary

The optimization-based attack method works by fixing the network parameters and taking the input of the model as the value to be updated. Depending on the output of the model, the main optimization objective is to reduce the value of the loss function. Good progress has been made in recent years.

The essence of the gradient-based attack method is the same as the optimization-based adversarial sample algorithm, which first calculates the gradient of the input data and then gradually updates the input data through the gradient according to the meaning of the loss function, which has a unique advantage in sample migration.

The boundary-based attack method solves the problem that the gradient-based method needs to set a hyperparameter learning rate in the process of gradient iteration, and perturbs the samples directly through the decision boundary.

The GAN-based attack method, on the other hand, links the adversarial attack with the generative adversarial network, and the sample generation has higher realism and good aggressiveness to all existing defenses, which helps us further optimize the deep learning model.

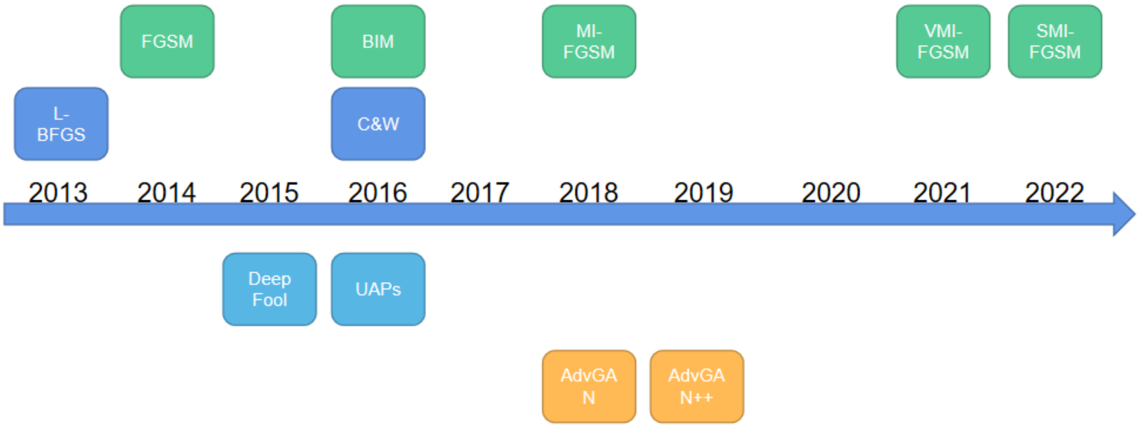

Figure 2. The development of white-box attacks.

As Figure 2 shows, White-box attacks have an earlier and deeper development, however, their nature of requiring a complete access model also makes them poorly used in the industry level.

4. Black-box attack

Black-box attack, i.e., an attack means to generate adversarial samples against a network model without being able to know the training process and parameters of the model [16].

Current white-box attacks have made considerable progress, but their attack conditions are demanding and require the attacker to have full access to the deep learning model, so a black-box attack that can perform an adversarial attack on this model only through the input and output provided by the application service is more relevant in practice.

4.1. Query-based attack methods

In a black box attack, the attacker can only use feedback information to carry out the attack, and depending on the feedback received by the target model, query-based attacks can be further classified into score-based and boundary-based attacks. Score-based attacks receive feedback from the target model in the form of probabilities or scores, while boundary-based attacks receive feedback in the form of hard labels.

4.1.1. Score-based attack methods. While white-box attacks have methods for attack sample generation based on the gradient generated by the model on the input samples, in black-box attacks the countermeasure samples can be generated by approximating the gradient of the model based on its output, i.e., based on the fraction. Chen et al. first used the pioneering zero-order optimization method (ZOO) to approximate the target model by valuing the first-order gradient and the second-order gradient through the symmetric quotient gradient to generate the adversarial sample [17]. And after obtaining the gradients, the adversarial samples are updated using a stochastic coordinate descent method.

ZOO can achieve similar results as the gradient-based white-box attack, but estimating the gradient requires extensive computation and multiple iterations, so this leads to a low efficiency of the ZOO attack.

AutoZOO, proposed by Tu et al., further optimizes the scheme based on ZOO by performing a zero-order optimization method based on autoencoder, which significantly improves the efficiency of the attack [18].

The average number of queries executed by black-box attacks is still high. The most effective attacks still require tens of thousands or more queries. Based on these problems, GUO et al. proposed Simple Black-box Adversarial Attacks (SimBA), a simple and efficient black-box sample attack method, which repeatedly selects a random direction from a pre-specified set of orthogonal search directions and uses the confidence level to check whether it points to or away from the decision boundary, based on which the image is randomly perturbed by adding or subtracting vectors [19]. Each update moves the image from stepwise to the decision boundary. The success rate of this approach is similar to that of state-of-the-art black-box attack algorithms, but the number of black-box queries is unprecedentedly low. SimBA is a new surprisingly powerful baseline for adversarial image attacks, and may have further applications in other areas in the future.

4.1.2. Decision-based attacks methods. Decision-based attacks in a stricter black-box environment, i.e., after the samples are input to the target model, only the categories (i.e., hard labels) assigned to the samples by the model are accessible, in a sense that decision-based attack methods are closer to the real scenario of the average attacker and therefore more difficult.

The Boundary Attack proposed by Brendel et al. is the first solution for a black-box attack when only the hard labels of the target model output are available [20]. First, the original image is initialized randomly, and each iteration generates a perturbation and makes the adversarial sample still within the valid range of the image at each perturbation step, and the distance between the perturbation and the adversarial sample and the original sample is proportional. The adversarial samples are generated by iterating continuously while keeping close to the original samples. And it also works well for the model after defensive distillation, but the drawback is that the average number of queries executed is too high and the convergence is poor.

TREMBA, proposed by Huang et al, makes improvements to address the above problems by using a pre-trained codec that enables it to generate effective interference with the target network in the low-dimensional embedding space, and then performs an effective search in the embedding space to generate adversarial samples to attack the unknown network [21]. This method reduces the number of visits and improves the success rate of the attack. And since TREMBA uses the global information of the source model to obtain the adversarial features that are insensitive to different models, it improves the migration of perturbations under different models.

The Triangle Attack that Wang et al. proposed achieves the first direct optimization of perturbations in frequency space using geometric information. Unlike most decision-based attacks, it does not need to perform gradient estimation at each iteration or restrict xadvt on the decision boundary, resulting in higher query efficiency and good performance in industrial applications [22].

4.2. Migration-based attack methods

Goodfellow et al. found that structurally similar neural network models were effective to some extent against the same adversarial sample attacks [16]. This implies that the adversarial samples generated using white-box attacks may also be effective against unknown models, i.e., the adversarial samples possess migration capabilities.

4.2.1. VMI-FGSM. Wang et al. did further research on various adversarial attacks based on FGSM for their mobility [11]. The method of variance-based adjustment of iterative gradient was proposed to improve the mobility of produced adversarial samples. The proposed VMI-FGSM performs better in migration to black-box attacks compared to all previous FGSM migrations.

4.2.2. MGAA. Based on the idea of meta-learning, Yuan et al. proposed Meta Gradient Adversarial Attack (MGAA) by randomly sampling multiple models from a collection of models to compose different tasks, and selecting multiple models to simulate a white-box attack and one model to simulate a black-box attack in each task, respectively [23].

MGAA narrows the gap between the gradient directions of white-box attacks and black-box attacks and can improve the mobility of the adversarial samples to the black-box setup.

4.3. Substitution-based attack methods

In traditional query-based black-box attack methods, either classification probability-based attacks or attacks based on hard labels returned by the target model, and many queries about the target model are required. The migration-based attack approach also requires a stand-in model for obtaining adversarial samples, followed by a migratory attack on the target model based on these adversarial samples. This still requires a lot of real data of the target model to train the stand-in model, so it is important to propose a method that can train the stand-in model without data.

4.3.1. DaST. Based on the above problems, Zhou et al. proposed Data-free Substitute Training (DaST), which uses a generative adversarial network GAN to generate synthetic samples to train the stand-in model, while the labels of the synthetic samples come from the target model [24].

The GAN-based generator randomly samples noise from the input space and synthesizes the data, and then feeds the synthesized data into the target model to obtain the output data. The stand-in model is trained using the generated input and output data pairs, and Figure 3 shows the principle of DaST.

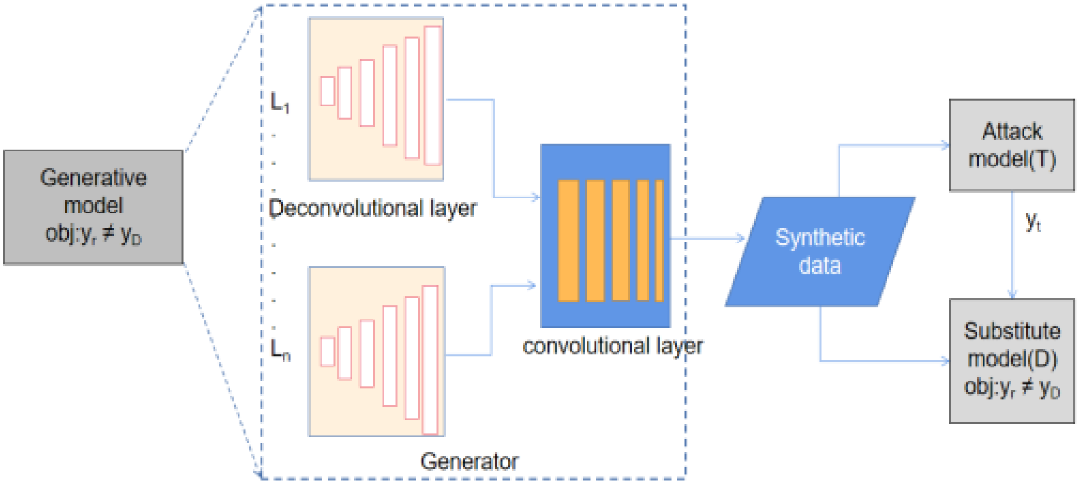

Figure 3. Principle of DaST.

DaST implements the first approach to train a stand-in model without any real data, so that an attacker can train a stand-in model to achieve an attack without real data. It also shows that current machine learning systems are still exposed to risks due to the lack of deep learning interpretability [1,25].

4.3.2. Delving into data. Delving into Data by Wang et al. proposed a new effective Diverse data generation module based on the generative alternative training paradigm by delving into the nature of the input "generated alternative training data", while introducing the adversarial Alternative training brings the target model boundary closer to the boundary of the alternative model [26]. The combination of the two modules results in a better agreement between the alternative model and the target model than DaST [24].

Delving into Data also demonstrates that substitution-based training with GAN-generated artificial data can yield better results than the real dataset, and the advantages are significant.

4.4. The single-pixel attack

The single-pixel attack, proposed by Su et al. analyzes a sample attack in a very limited scenario where only one pixel can be adjusted at a time [27]. A single-pixel black-box attack approach is proposed in the paper using differential evolution.

This method demonstrates that many data points may be located near the decision boundary and even very small perturbations can accumulate over the values of multiple dimensions and lead to large changes in the output. However, this method is less effective the larger the image size is, and a large number of iterations is required if a better solution is to be found.

4.5. Summary

The disadvantages of query-based attack methods are the high number of queries, easy to be defended and detected by the target model, and low attack efficiency. Therefore, the improvement goal of query-based black box attacks is to reduce the number of queries. Among them, boundary-based attacks are closer to real scenarios and do not need to rely on alternative models, but they also bring the following challenges: classification labels are insensitive to gradient-based perturbations; input samples to the target model can only obtain hard labels, so they lead to discontinuous objective functions and high optimization difficulties. The score-based attack is more efficient but requires obtaining classification confidence and has a narrow application.

The migration-based attack method does not require querying the target model but generally requires a dataset to train the alternative model, and the success rate of the attack is not high. The main improvement idea is to narrow the gap between the alternative model and the actual model to improve the migration of the samples.

The substitution-based attack method proposes a new and more efficient idea to achieve better results on noise-based generated virtual data, which deserves further research and consideration on deep learning models and counterattacks.

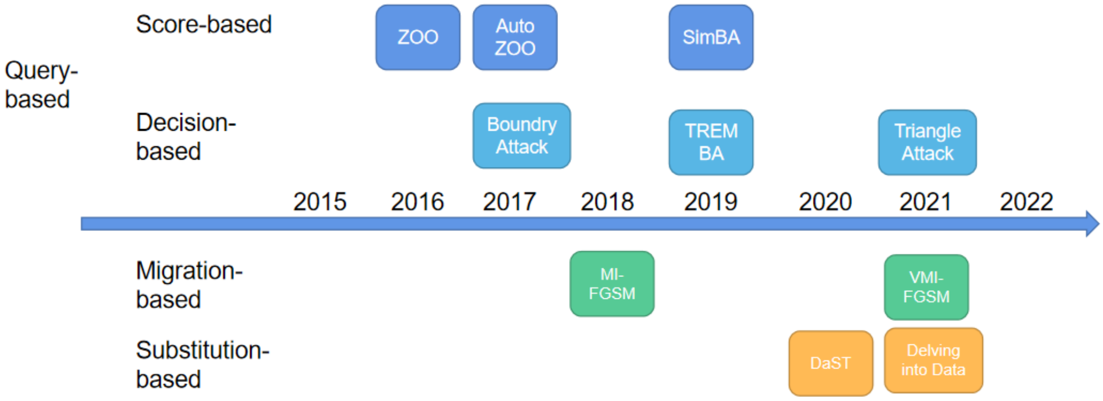

Figure 4. Development of Black-box attacks.

As Figure 4 shows, in recent years, black-box attacks have made considerable progress on many levels and are believed to have better prospects in the future due to their lower model-awareness requirements and their higher portability.

5. Conclusion

This paper summarizes the basic principles and classification methods of deep neural network adversarial attacks, briefly introduces related concepts and definitions, and selects several important and academic frontier areas to interpret and analyze. The development of adversarial attacks is reviewed. The early research on adversarial attacks mainly focuses on their efficiency. Then the mobility, the effectiveness of adversarial defense, and the use and deployment in other low-control conditions are extended and studied.

The phenomenon of adversarial attacks has gained a lot attention recently, however, adversarial attacks against deep learning still deserve attention until the interpretability of deep learning is fully unveiled. In this paper, only the adversarial sample attack method and its related definitions are introduced, and further research can focus on target detection and sample robustness. It is expected that the work in this paper can provide useful references for future researchers and make a little contribution to the learning and research of sample-based adversarial attacks on deep neural networks.

References

[1] Ren, K., Zheng, T. and Qin, Z., et al. (2020). Adversarial Attacks and Defenses in Deep Learning. Engineering 6(3), 15.

[2] Moosavi-Dezfooli, S. M., Fawzi, A. and Fawzi, O., et al. (2017). Universal adversarial perturbations. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

[3] Liu, Y., Chen, X. and Liu, C., et al. (2016). Delving into Transferable Adversarial Examples and Black-box Attacks. DOI:10.48550/arXiv.1611.02770.

[4] Szegedy, C., Zaremba, W. and Sutskever, I., et al. (2013). Intriguing properties of neural networks. Computer Vision and Pattern Recognition (cs.CV). Machine Learning (cs.LG). Neural and Evolutionary Computing (cs.NE). https://doi.org/10.48550/arXiv.1312.6199.

[5] Papernot, N., Mcdaniel, P. and Wu, X., et al. (2016). Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. 2016 IEEE Symposium on Security and Privacy (SP). IEEE.

[6] Carlini, N., Mishra, P. and Vaidya, T., et al. (2016). Hidden voice commands. In 25th USENIX Security Symposium (USENIX Security 16), Austin, TX.

[7] Goodfellow, I., Shlens, J. and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In: Proceedings of 2015 International Conference on Learning Representations (ICLR Poster). San Diego, CA, USA.

[8] Kurakin, A., Goodfellow, I. and Bengio, S., et al. (2017). Adversarial examples in the physical world. In: Proceedings of 2017 International Conference on Learning Representations (ICLR). Toulon, France, 1-14.

[9] Dong, Y., Liao, F. and Pang, T., et al. (2018). Boosting adversarial attacks with momentum. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9185-9193.

[10] Wang, X. and He, K. (2021). Enhancing the transferability of adversarial attacks through variance tuning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1924-1933.

[11] Wang, G., Yan, H. and Wei, X. (2022). Enhancing Transferability of Adversarial Examples with Spatial Momentum. Accepted as Oral by 5-th Chinese Conference on Pattern Recognition and Computer Vision, PRCV 2022. https://doi.org/10.48550/arXiv.2203.13479.

[12] Moosavi-Dezfooli, S., Fawzi, A. and Frossard, P., et al. (2016). DeepFool: A simple and accurate method to fool deep neural network. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2574-2582. DOI:10.1109/CVPR.2016.282.

[13] Moosavi-Dezfooli, S. M., Fawzi, A. and Fawzi, O., et al. (2017). Universal adversarial perturbations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1765-1773.

[14] Xiao, C., Li, B. and Zhu, J. Y., et al. (2018). Generating Adversarial Examples with Adversarial Networks. Cryptography and Security (cs.CR). Computer Vision and Pattern Recognition (cs.CV). Machine Learning (stat.ML). https://doi.org/10.48550/arXiv.1801.02610.

[15] Mangla, P., Jandial, S. and Varshney, S., et al. (2019). AdvGAN++: Harnessing latent layers for adversary generation. Accepted at Neural Architects Workshop, ICCV 2019. Computer Vision and Pattern Recognition (cs.CV). Machine Learning (cs.LG). https://doi.org/10.48550/arXiv.1908.00706.

[16] Papernot, N., Mcdaniel, P. and Goodfellow, I., et al. (2016). Practical Black-Box Attacks against Machine Learning. ACM.

[17] Chen, P. Y., Zhang, H. and Sharma, Y., et al. (2017). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 15-26.

[18] Tu, C. C., Ting, P. and Chen, P. Y., et al. (2018). AutoZOOM: Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks. Computer Vision and

Pattern Recognition (cs.CV). Cryptography and Security (cs.CR). Machine Learning (stat.ML). https://doi.org/10.48550/arXiv.1805.11770.

[19] Guo, C., Gardner, J. R. and You, Y., et al. (2019). Simple Black-box Adversarial Attacks. Machine Learning (cs.LG). Cryptography and Security (cs.CR). Machine Learning (stat.ML). https://doi.org/10.48550/arXiv.1905.07121.

[20] Brendel, W., Rauber, J. and Bethge, M. (2017). Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. Machine Learning (stat.ML). Cryptography and Security (cs.CR). Computer Vision and Pattern Recognition (cs.CV). Machine Learning (cs.LG). Neural and Evolutionary Computing (cs.NE). https://doi.org/10.48550/arXiv.1712.04248.

[21] Huang, Z. and Zhang, T. (2019). Black-Box Adversarial Attack with Transferable Model-based Embedding. Machine Learning (cs.LG). Machine Learning (stat.ML). https://doi.org/10.48550/arXiv.1911.07140.

[22] Wang, X., Zhang, Z. and Tong, K., et al. (2021). Triangle Attack: A Query-efficient Decision-based Adversarial Attack. Computer Vision and Pattern Recognition (cs.CV). https://doi.org/10.48550/arXiv.2112.06569.

[23] Yuan, Z., Zhang, J. and Jia, Y., et al. (2021). Meta Gradient Adversarial Attack. Computer Vision and Pattern Recognition (cs.CV). https://doi.org/10.48550/arXiv.2108.04204.

[24] Zhou, M., Wu, J. and Liu, Y., et al. (2020). DaST: Data-free Substitute Training for Adversarial Attacks. Cryptography and Security (cs.CR). Computer Vision and Pattern Recognition (cs.CV). Machine Learning (cs.LG). https://doi.org/10.48550/arXiv.2003.12703.

[25] Ilyas, A., Santurkar, S. and Tsipras, D., et al. (2019). Adversarial Examples Are Not Bugs, They Are Features. Machine Learning (stat.ML). Cryptography and Security (cs.CR). Computer Vision and Pattern Recognition (cs.CV). Machine Learning (cs.LG). https://doi.org/10.48550/arXiv.1905.02175.

[26] Wang, W., Yin, B. and Yao, T., et al. (2021). Delving into Data: Effectively Substitute Training for Black-box Attack. Computer Vision and Pattern Recognition (cs.CV). https://doi.org/10.48550/arXiv.2104.12378.

[27] Su, J., Vargas, D. V. and Kouichi, S. (2017). One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation 23(5), 828-841. https://doi.org/10.48550/arXiv.1710.08864.