Studies Advanced in Image Recognition based on Adversarial Learning

Jiaqi Liao[1*]

[1] Department of Computer Science, Faculty of Science & Engineering, University of Liverpool, Liverpool L69 3BX, United Kingdom
*Corresponding author:J.Liao12@student.liverpool.ac.uk

Abstract. In recent years, adversarial learning has gradually attracted a lot of research interest, which aims to understand the attack behavior and design various algorithms that can resist the attack. The design of adversarial learning algorithms mostly revolves around the generation of adversarial examples, which refer to samples that are carefully crafted for these recognition tasks to confuse and mislead detection tasks. Adversarial learning finds applications in various domains including medical care, finance, security, and autonomous driving, demonstrating promising prospects. Taking the classical image recognition task as an example, this paper provides a detailed overview of recent developments in adversarial learning. Specifically, two main frameworks and corresponding representative algorithms of adversarial learning are introduced, including their design ideas, key steps, advantages, and disadvantages.. Then, quantitative results of different classification algorithms on common datasets are analyzed and compared. The article concludes by summarizing the difficulties in balancing accuracy and robustness, parameter settings, algorithm selection, and prospects for the future development of adversarial learning, which can provide some new insights for the research field of adversarial learning.

Keywords: Image recognition, Adversarial learning, Deep learning.

1 Introduction

With the rapid advancements of computer vision and machine learning, image classification has emerged as a crucial task in the field of artificial intelligence. The goal of image classification is to accurately assign input images to predefined categories, and it has wide applications in various domains such as face recognition, object detection, and autonomous driving. However, researchers have increasingly recognized the vulnerability and susceptibility to attacks exhibited by models. Real-world data and environments are often filled with noise and interference. For example, input data may be maliciously modified or randomly perturbed by attackers, leading to unacceptable errors in traditional machine learning models.

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_97

To address the above problems, image recognition based on adversarial learning has attracted extensive attention in recent years. As a promising solution to model robustness and security, adversarial learning mainly focuses on exploring and analyzing the behavior and performance of models in the face of targeted attacks. Adversarial learning typically involves two components: adversarial samples and the target classifier model. The perturbations in adversarial samples are indistinguishable to humans but can deceive the model. The target classifier model is trained to correctly recognize the original images and resist misclassification when presented with adversarial samples, thereby enhancing the model's ability to differentiate between real and adversarial images. This helps models better cope with interference and attacks, enhancing their robustness and reliability, and enabling them to better adapt to complex environments in real-world scenarios.

Adversarial learning methods can be divided into Generative Adversarial Networks (GANs) and Adversarial Training. GANs are an important adversarial learning method, whose main idea is to generate realistic samples by adversarial training between generators and discriminators. The generator is responsible for generating the faked samples, while the discriminator is used to distinguish the real samples from the generated ones. Through a feedback mechanism, the generators and discriminators compete with each other and improve upon each other, with the result that the generators are able to generate realistic samples that are indistinguishable from the real samples. Another common adversarial learning method is Adversarial Training, whose main idea is to inject adversarial perturbations or introduce hostile samples into the training data in order to improve the robustness of the model to small perturbations in the input. In this way, the model can have better generalization and defense capabilities.

In the context of image classification tasks, adversarial learning specifically investigates attack methods and defense strategies that can manipulate image inputs to deceive models. Focusing on the above two categories of adversarial learning frameworks, this paper provides an overview of recent research advances and related methods in adversarial learning for image classification. In detail, the evolution of the image classification task and its applications and challenges are reviewed first. Then, the concepts and fundamentals of adversarial learning will be introduced, including Generative Adversarial Networks (GAN) and adversarial attacks on samples. A systematic overview of currently used adversarial attack methods and defense mechanisms is further provided, which is followed by an analysis of their strengths and weaknesses. Finally, the potential directions and challenges for future research on adversarial learning for image classification are discussed, with a view to providing insights and guidance for further research in the field.

2 Method

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) [1] are a type of deep learning model that operates on the principle of training two competing neural network models: the

generator and the discriminator. The objective of the generator is to learn the skill of generating authentic-looking samples from random noise, whereas the discriminator's task is to distinguish between samples produced by the generator and genuine samples. Through iterative training, the generator and discriminator engage in a mutual adversarial learning process, gradually improving the generator's ability to produce more realistic samples while enhancing the discriminator's accuracy.

When GANs are used to enhance the robustness of classifiers, the classifier serves as the discriminator. It strives to accurately differentiate between genuine samples and adversarial samples generated by the generator in collaboration with the attack algorithm. The iterative adversarial training process enables the classifier to learn robust representations against adversarial perturbations, thereby improving the classifier's performance and robustness. This approach has been empirically validated in the Generative Adversarial Trainer (GAT) [2] proposed by Hyeungill et al. Compared to other methods, GANs offer greater flexibility and creativity in generating adversarial samples, which leads to superior robustness in classifiers. However, GAN training is computationally intensive and requires more resources and time. On the other hand, conventional adversarial training algorithms are simpler and more direct but may exhibit relatively weaker performance in terms of the quality of generated adversarial samples and their effectiveness in attacks.

Pouya et al. introduced Defense-GAN [3] in 2018 as a technique that employs generative models to safeguard classifiers from adversarial attacks. The primary goal of Defense-GAN is to learn the distribution of original, undisturbed images through training. During inference, it generates an output similar to the input image but without any adversarial modifications, which is then fed into the classifier. One significant advantage of this approach is its compatibility with various classification models, as it does not require any alterations to the structure or training process of the classifier. It can function as an additional defense mechanism against any attack, as it doesn't rely on prior knowledge of adversarial example generation. Experimental findings consistently demonstrate that Defense-GAN effectively combats different attack methods and enhances existing defense strategies. However, it should be noted that Defense-GAN may encounter overfitting problems during adversarial training. This means that the model may only be capable of adapting to specific types of adversarial attacks and could lack robustness against unknown attacks.

2.2 Adversarial Training

Adversarial training is a method to enhance model robustness and defense capability, whose idea is to introduce adversarial samples or perturbations during the training process. Firstly, adversarial samples or perturbations are constructed using certain attack methods. The model is then trained using the attack dataset, giving it the ability to cope with the difficulties presented by both the original and adversarial samples. The model steadily increases its resilience to adversarial attacks by learning robust classification against adversarial inputs. The Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini-Wagner (CW) attack are three distinctive and popular attack methods that are specifically highlighted in this

study. Other methods include One Pixel Attack, Deepfool, Jacobian-based Saliency Map Attack (JSMA), among others.

Fast Gradient Sign Method (FGSM) [4] is one of the most fundamental and widely used attack methods, upon which many other methods are based or improved. FGSM attack training introduces adversarial perturbations during the training process, enabling the model to better resist small perturbations in the input. FGSM constructs adversarial samples by computing the gradient sign of the input sample. Specifically, given an input sample x and its corresponding label y, the gradient of the loss function with respect to the input sample is computed, and the resulting sign is multiplied by a small perturbation ε to generate the adversarial perturbation δ. Finally, the original input sample is combined with the adversarial perturbation to produce the adversarial sample $(x' = x + \delta)$. The constructed adversarial sample appears very similar to the original sample in appearance, but it can deceive the model into producing incorrect classification results. The advantages of FGSM, as the most basic attack method, lie in its simplicity and efficiency. To create the adversarial sample, all that required is to compute the gradient sign of the input sample and then apply a perturbation. However, its drawback is that it often generates only slight perturbations that maintain visual similarity with the original sample. Wong et al. [5] proposed using weaker and cheaper adversarial ensembles to train robust models, achieving the same effectiveness as PGD attack training.

Projected Gradient Descent (PGD) [6] is an iterative adversarial attack method that generates adversarial samples by iteratively applying gradient ascent on the input samples to deceive the model. The main idea of PGD attack training is to construct more challenging adversarial samples by repeatedly applying gradient ascent to maximize the loss function during the training process. In each training step of PGD attack training, the current model is used to forward propagate the selected adversarial samples, and the loss function is computed. Then, the gradient of the loss function with respect to the input samples is calculated, and a projection operation is applied to constrain the updates within a step size range, epsilon, to obtain the next adversarial sample. This process is repeated for multiple iterations, generating updated adversarial samples at each iteration. As a result, the adversarial samples gradually approach the decision boundary of the model, posing a greater challenge to the model. The attack principle of PGD, i.e. projected gradient descent over a negative loss is shown in Equation (1).

$$x_{t+1} = \Pi_{x+S}\left(x_t + \alpha \cdot sign\left(\nabla_{x_t} J(\theta, x_t, y)\right)\right)$$
$$S = x' \parallel x - x'|_\infty \leq \epsilon \tag{1}$$

where x_t denotes the input sample at the t^{th} iteration, α denotes the step size (or learning rate), S denotes the set of constraints with L_∞ as epsilon, and $x + s(x')$ denotes the projection of x' to the nearest point in the set of constraints S. In this way, after each iteration, the adversarial samples are projected back into the constraint set to ensure that they satisfy the constraints of the L_∞ paradigm.

PGD attack is the strongest first-order attack [6] and is regarded as a strong and effective adversarial attack strategy, as it can overcome many defense measures. The advantage of training with a PGD attack lies in its iterative nature, which generates more challenging adversarial samples. Compared to single-step attack methods, PGD attacks can comprehensively explore the model's vulnerabilities and provide stronger defense capabilities. However, the main drawback of PGD attack training is its increased demand for computational resources and time, as each sample requires multiple iterations to generate adversarial samples, which may be impractical for certain large-scale datasets.

Carlini and Wagner (CW) [7] proposed three adversarial attack methods in 2017 to effectively attack Defensive Distillation networks. In contrast to FGSM and PGD, experimental results have shown that the C&W attack method can effectively bypass most existing defense mechanisms. The primary principle behind CW attack training is to generate adversarial examples by optimizing an objective function, aiming to deceive the model into misclassifying while minimizing the magnitude of adversarial perturbations. Specifically, CW solves an optimization problem to generate adversarial examples, where the objective function consists of an adversarial loss to induce misclassification and a regularization term to control the size of adversarial perturbations. CW attack training possesses several advantages over other adversarial training methods. Firstly, it is capable of generating more challenging adversarial examples by simultaneously considering misclassification and the magnitude of adversarial perturbations through objective function optimization. Secondly, CW can adapt to different threat models and generate effective adversarial examples even without knowledge of the attack algorithm. Additionally, CW allows for balancing the importance of misclassification and perturbation by adjusting the parameters of the optimization problem to meet specific application requirements. The drawbacks of CW include relatively higher computational complexity as it requires solving an optimization problem to generate adversarial examples. Furthermore, CW may impact the training and inference time of the model as it involves an additional optimization process for generating adversarial examples.

3 Experiment

3.1 Dataset

Common image classification datasets mainly include MNIST [8], and CIFAR-10 [9]. MNIST is a classic handwritten digital image classification dataset containing 60,000 handwritten digital images for training and 10,000 handwritten digital images for testing. These images contain numbers from 0 to 9. Each image has a size of 28x28 pixels and is represented as a grey-scale image. The MNIST dataset is typically used as a benchmark test for machine learning algorithms and models in tasks like classification.

The CIFAR-10 dataset, which is frequently used for image classification, consists of 60,000 color images divided into 10 categories, which include different vehicles and animals. The size of every image is 32×32 pixels. the CIFAR-10

dataset has a high diversity and complexity and is closer to image classification tasks in real scenarios than the MNIST dataset. This makes CIFAR-10 a common benchmark for evaluating the performance of various image classification algorithms and models. In addition, the CIFAR-10 dataset is one of the commonly used datasets for conducting adversarial attacks and robustness research.

3.2 Quantitative comparison

To quantitatively evaluate the performance of different methods, experiments were carried out on different datasets. The results of adversarial attack experiments conducted on the MNIST dataset can be observed in Table 1. Each attack in the table adopts different attack methods, perturbation sizes, and iteration counts. The source include the model itself for white-box attacks (A), an independent network (A') for black-box attacks, and an framework from [10] (B). The most effective attacks are indicated in bold for each attack model.

As demonstrated in Table 1, PGD succeeds in the best performance on both A and A' datasets. Particularly, the results for attacks on the network itself (A) with different parameters are superior to FGSM and PGD. The performances of FGSM and CW algorithms are comparable to each other.

Table 1. Performance of various methods on the MNIST dataset

Method	Source	Steps	Restarts	Accuracy
Original	-	-	-	98.8%
FGSM	A	-	-	95.6%
CW	A	40	1	94.0%
PGD	A	40	1	93.2%
PGD	A	100	1	91.8%
PGD	A	40	20	90.4%
PGD	A	100	20	**89.3%**
FGSM	A'	-	-	96.8%
CW	A'	40	1	97.0%
PGD	A'	40	1	96.0%
PGD	A'	100	20	**95.7%**
FGSM	B	-	-	**95.4%**
PGD	B	40	1	96.4%

Table 2 provides the results of attacks made using the same configuration on the CIFAR-10 dataset. The source used for the attacks include the network itself for white-box attacks (A), an independent network (A') for black-box attacks, and a

network trained on the original dataset (Anat). The most effective attacks are indicated in bold for each attack model. For the more challenging real-world dataset, the attack algorithms demonstrate a stronger impact. This is demonstrated through white-box attacks, where the accuracy on the adversarial dataset decreases by up to 41.5% in comparison to the classifier's accuracy on the original dataset. Among white-box attacks, even after parameter tuning, PGD remains the strongest attack method. However, CW attacks are also highly effective, while FGSM is relatively weaker. In the case of black-box attacks (A'), PGD is the strongest, followed by FGSM, and CW is the weakest. When attacking the network Anat , which has undergone adversarial training, the difference in accuracy compared to the classifier network on the original data is minimal, indicating that adversarially trained models exhibit stronger robustness compared to models without adversarial training.

Table 2. Performance of various methods on the CIFAR-10 dataset

Method	Source	Steps	Accuracy
Original	-	-	87.3%
FGSM	A	-	56.1%
CW	A	30	46.8%
PGD	A	20	**45.8%**
PGD	A	7	50.0%
FGSM	A'	-	67.0%
CW	A'	30	78.7%
PGD	A'	7	**64.2%**
FGSM	$Anat$	-	85.6%
PGD	$Anat$	7	86.0%

The Table 3 shows the average classification accuracy using different defense strategies under different attack conditions. It can be found that Defense-GAN significantly outperforms MagNet and adversarial training. This indicates that GAN has a significant advantage in terms of effectiveness in improving the robustness of the model.

Table 3. Average classification accuracy using different defense strategies

Attack	No Attack	No Defence	**Defence-GAN**	MagNet	Adv. Training
FGSM	0.986	0.608	**0.978**	0.133	0.557
RAND+FGSM	0.986	0.087	**0.974**	0.132	0.589
CW	0.986	0.083	**0.969**	0.030	0.100

4 Discussion

Accuracy and robustness must be traded off in adversarial attacks. Adversarial training can potentially decrease the accuracy of a model on normal inputs while increasing its robustness against adversarial attacks. This trade-off needs to be carefully considered to ensure that the model remains useful for its intended task. Further research is needed to explore various adversarial learning methods.

In adversarial learning, parameter tuning is a challenging task that significantly impacts the final results. Whether it is adversarial training or training based on Generative Adversarial Networks (GANs), selecting appropriate parameters is crucial for ensuring the model's robustness. However, there is no universal standard for determining these parameters, as different datasets, models, and attack scenarios may require different settings. Parameter tuning often involves a combination of experimentation and expertise, relying on human intuition and domain knowledge, necessitating further investigation.

Choosing suitable algorithms in adversarial learning is highly challenging due to the diversity and complexity of attack and defense methods. One difficulty lies in the diversity of algorithms. In adversarial learning, there exist various attack and defense methods, each with its specific advantages and limitations. For example, when generating attack datasets, one can choose between fast and efficient first-order methods like Fast Gradient Sign Method (FGSM) or more complex but powerful iterative methods like Projected Gradient Descent (PGD). Regarding defensive training, different methods such as adversarial training, Generative Adversarial Networks (GANs), among others, can be employed. Selecting the appropriate algorithm requires considering the complexity of the problem, dataset characteristics, model performance requirements, available computational resources, and other factors. Another challenge is algorithm evaluation and comparison. Evaluating algorithms in adversarial learning involves multiple metrics such as attack success rate, defense success rate, and robustness indicators, which may conflict with each other, such as the trade-off between robustness and accuracy. Therefore, it is necessary to consider multiple metrics to comprehensively evaluate the algorithm's overall performance and make trade-offs in practical applications.

With the continuous development and research in the field of adversarial learning, we can expect more guiding principles and methods to help select appropriate algorithms and improve model performance and robustness. Further exploration of automated methods for algorithm selection and optimization can reduce reliance on manual parameter tuning and discover better parameter configurations and algorithm combinations. Additionally, the application of adversarial learning extends beyond image recognition, and exploring algorithm transferability and generalization is also worth investigating.

5 Conclusion

This paper provides a concise overview of adversarial learning in image classification. Two primary approaches to adversarial learning and their respective typical algorithms are introduced, followed by quantitative data analysis and comparisons of different algorithms applied to common datasets and base classifiers. The paper concludes by summarizing the current challenges in adversarial learning, including the balance between accuracy and robustness, parameter settings, and algorithm selection, and provides prospects for the future development of adversarial learning. This paper is aimed to provide researchers with an introduction to adversarial learning in image classification and promote further advancements in this field. Adversarial learning has the potential to enhance model security and improve the reliability of image classification tasks, thus enabling better real-world applications.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y.: Generative adversarial networks. In: Communications of the ACM, 63(11): pp. 139-144. (2020).
2. Lee, H., Han, S., & Lee, J.: Generative adversarial trainer: Defense to adversarial perturbations with gan. arXiv preprint arXiv:1705.03387 (2017).
3. Samangouei, P., Kabkab, M., & Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. arXiv preprint arXiv:1805.06605 (2018).
4. Goodfellow, I. J., Shlens, J., & Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014).
5. Wong, E., Rice, L., & Kolter, J. Z.: Fast is better than free: Revisiting adversarial training. arXiv preprint arXiv:2001.03994 (2020).
6. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017).
7. Carlini, N., & Wagner, D.: Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM workshop on artificial intelligence and security. (2017).
8. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE, 86(11): pp. 2278-2324. (1998).
9. Krizhevsky, A., & Hinton, G.: Learning multiple layers of features from tiny images. (2009).
10. Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P.: The space of transferable adversarial examples. arXiv preprint arXiv:1704.03453 (2017).