Research on Attack and Defence of Face Recognition Based on Adversarial Learning

Chunhao Zhang

School of Intelligent Software and Engineering, Nanjing University, Suzhou, Jiangsu, 215004, China

Abstract. The rapid advancement of face recognition technology brings inevitable threats from various attacks, significantly reducing model accuracy and posing serious security risks. Enhancing defence mechanisms through algorithms is crucial to mitigate these impacts. However, these methods have limitations. New or complex attacks quickly diminish their effectiveness as attackers upgrade their techniques. Some defence techniques, while boosting model resilience, can adversely affect performance. This paper focuses on adversarial learning for attack and defence in face recognition, demonstrating the vulnerability of general models to attacks and giving a defence method that has some effect, while highlighting the limitations of current defences against complex attacks. Implementing adversarial learning reveals stark vulnerabilities: the constant tuning of parameters in the FGSM attack method reduces the model accuracy to 39.00%, illustrating its aggressiveness. In contrast, the adoption of PGD increases the sample prediction rate of the attacked prediction errors to 26.00%, demonstrating a viable defence strategy to bolster model robustness against such threats.

1 Introduction

Face recognition technology in recent years, it can efficiently and accurately carry out identification and is widely used in security, finance and other fields to improve safety and convenience. Its development has undergone a transformation from traditional feature extraction to deep learning models, especially convolutional neural network (CNN), which automatically extracts face features through multi-layer convolution and pooling to achieve high-precision identification[1]. However, face recognition technology also encounters several security challenges, such as adversarial attacks, privacy issues, data bias, and model generalisation capabilities. Among them, adversarial attacks refer to the attackers adding tiny perturbations to the image that are hard to detect with the naked eye through careful design, misleading the face recognition system and seriously threatening the security and reliability of the face recognition system[2]. In the face of these challenges, Adversarial Learning was born. It has its origins in the early difficulties encountered in the practical application of machine learning and plays a key role in the field of face recognition. Attackers use the principles of adversarial learning to create confusion, and defenders use this technique to counteract attacks and increase the robustness of face recognition

221900068@smail.nju.edu.cn

systems[3]. With the deeper application of AI in critical areas, adversarial learning has gained attention for its ability to improve model safety and robustness.

There are several defence methods against face recognition attacks based on adversarial attacks. Adversarial training adds adversarial samples to the training data, and adds perturbations to simulate the attack each time training is performed, which enhances the robustness of the model to perturbations and improves the defence capability. However, it requires a large number of computational resources, takes a long time to train, is only effective against specific attacks (e.g., FGSM), is ineffective against new types of attacks, and is prone to overfitting, which affects the accuracy of the model[3, 4]. Feature compression uses methods such as PCA and self-encoder to reduce the dimensionality of the input data and reduce the perturbation space for counter-attacks, making the compressed features more difficult to be perturbed and enhancing the model's defence capability[5, 6]. However, the method may lose useful information, affecting accuracy, and has limited effect on high-dimensional adversarial samples, poor adaptability, and difficulty in dealing with various types of attacks. Input preprocessing reduces adversarial perturbations and improves system robustness through techniques such as denoising, image cropping and blurring[7, 8]. However, this can lead to loss of image details, affecting recognition accuracy, and simple pre-processing is difficult to cope with evolving and complex attacks[9]. While there are advantages and disadvantages to each of the existing defence methods, there is no perfect solution for different types of attacks.

This paper aims to test the aggressiveness of the FGSM attack and the defensiveness of the PGD adversarial learning method using the LFW dataset under the VGG16 model. Based on this dataset and model, this paper will demonstrate the impact of FGSM attack on the model performance through a series of experiments, including the changes of confrontation accuracy and attack success rate under different perturbation strengths, and at the same time elaborate the performance of the PGD adversarial learning method in defending against the attack, so as to compare the effect of attack and defence and analyse the essence of attack and defence in depth, and to reveal the inherent limitations of the defence in the case of complex attacks, thus providing a new perspective for the understanding of the essence of attack and defence. It provides a new perspective for the in-depth understanding of the nature of attack and defence.

2. Data and Methodology

2.1 Data sources

Fig. 1 Example of some images from the LFW dataset(Photocredit : Original)

Labeled Faces in the Wild (LFW) dataset is derived from real scenes, with more than 13,000 images, covering 5,749 different characters, with a wide range of characters' races, genders, and ages. The images vary greatly in posture, expression, lighting, etc., and are labelled in detail, which helps to improve the robustness and generalization ability of the model, and are widely used for face recognition algorithm evaluation[10]. Therefore, the LFW dataset is selected for this paper. Among them, Fig. 1 shows some sample images of the LFW dataset.

2.2 training model：VGG16

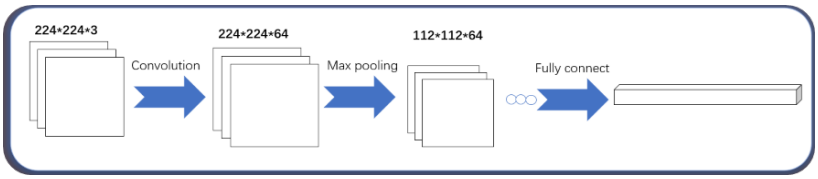


Fig. 2 shows part of the architecture of VGG16(Photocredit : Original)

As shown in Fig. 2, the VGG16 architecture consists of an input layer, a convolutional layer, a pooling layer, a fully connected layer and an output layer. The input layer accepts 224x224x3 RGB images, the convolutional layer uses 13 3x3 convolutional kernels divided into multiple convolutional blocks, each block is followed by a 2x2 maximum pooling layer, and the convolutional layer progressively extracts the image features and enhances the network representation. Finally, enter three fully connected layers, the last layer outputs the probability distribution of the category[11]. The advantage of VGG16 is its simple structure and the use of small 3x3 convolutional kernel stacked deep networks to effectively improve feature extraction. The deep architecture captures complex image features, is suitable for migration learning, and performs superiorly especially when the amount of data is insufficient. Its modular design is easy to extend and modify, and it shows strong robustness in a variety of computer vision tasks, with high efficiency and accuracy despite the large number of parameters[12].

2.3 Adversarial attack method

Fast Gradient Sign Method (FGSM) is a single-step adversarial attack method that quickly generates adversarial samples by computing the gradient sign of the model loss function on the input data and adding controlled perturbations (with magnitude controlled by ε) along the direction that maximizes the classification error. Its core formulation is equation (1) above. This method is computationally efficient and can significantly reduce model accuracy with little change in human eye perception, and is commonly used to expose the vulnerability of neural networks and assess model robustness, as well as to provide a base attack for adversarial training[4, 13].

2.4 Methods of defence

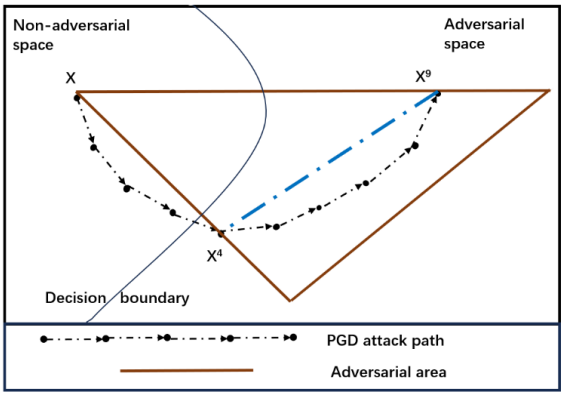
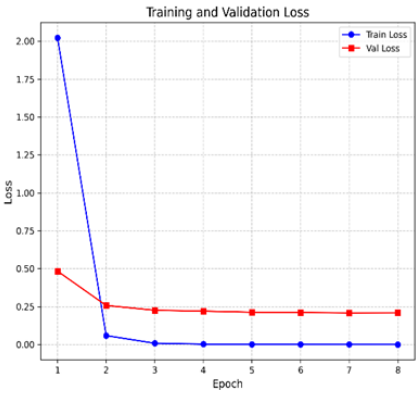


Fig. 3 PGD transfers samples from non-adversarial space to adversarial space by an iterative process(Photocredit : Original)

In this paper, input preprocessing is used to reduce the effect of anti-perturbation by modifying the input data (e.g. denoising, image enhancement, etc.)[9, 14]. This approach is simple and efficient, but it is usually not fully robust against sophisticated attacks. Adopting adversarial learning, adversarial samples are added to the training set during the training process, so that the model can improve the accuracy and enhance the robustness of the adversarial samples in an adversarial environment, and the common methods are adversarial training[15]. In this paper, Projected Gradient Descent (PGD) is used. Fig. 3 shows the process of PGD to transfer samples from non-adversarial space to adversarial space by an iterative way, which calculates the gradient of the perturbation to the loss function and updates the perturbation along the direction of the gradient to maximize the loss, and generates the adversarial samples by updating the perturbation through several iterations. It has the advantage of high flexibility to adjust the perturbation size, step size and number of iterations to enhance the attack effect[16].

3 Results

3.1 model training



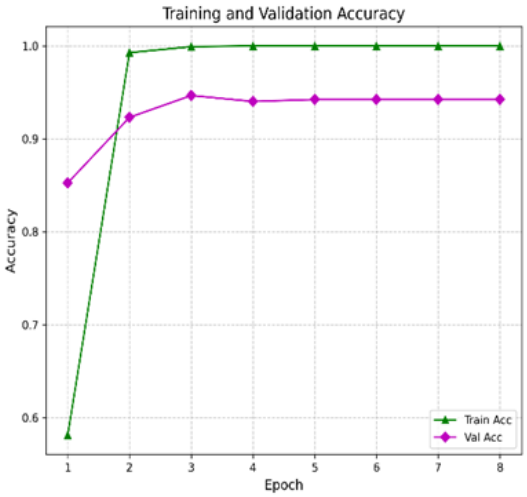


Fig. 4 Comparison of training and validation accuracies and loss values with the variation of Epoch(Photocredit : Original)

Fig. 4 demonstrates the model's accuracy and loss values change with Epoch during training and validation, the training loss and validation loss both decrease rapidly, and the training accuracy and validation accuracy increase rapidly and maintain at a high level. The training loss decreases from 2.0232 in the first round to a very low value in the late stage, and the training accuracy rapidly increases from 58.15% to stabilise at 100%; the validation loss gradually decreases from 0.4839 to a certain range of fluctuation, and the validation accuracy increases from 85.26% and stabilises at 94.23% in the late stage. The above data shows that it shows good accuracy on both training and test data, and the accuracy of the test set is close to that of the training set, which means that the model has good generalisation ability and is capable of performing face recognition tasks.

3.2 Attack studies

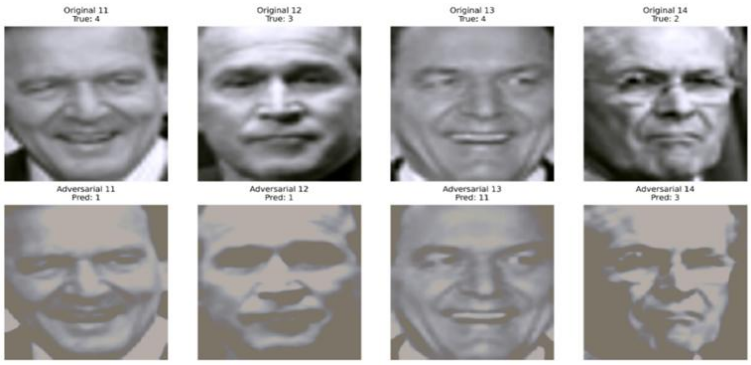


Fig. 5 Comparison between original image and generated adversarial image(Photocredit : Original)

Fig. 5 shows some of the attacked images and the original image, which shows that after the FGSM attack, the human eye is still able to distinguish different faces, however, the face recognition model is no longer able to recognise them, proving the vulnerability of the model to the attack.

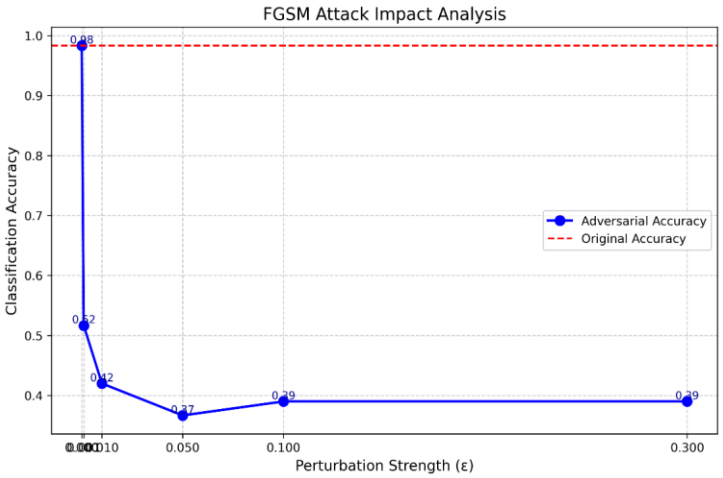


Fig. 6 Demonstrates the effect of FGSM attack parameter variations on model accuracy(Photocredit : Original)

Table 1. Generating Adversarial Samples Using FGSM Attacks

No	Perturbation strength (ε)	Original accuracy	Confrontation accuracy	Success rate of attack
2	0.01	98.33%	42.00%	57.29%
3	0.05	98.33%	36.67%	62.71%
4	0.1	98.33%	39.00%	60.34%
5	0.3	98.33%	39.00%	60.34%

Table 1 and Fig. 6 demonstrate the change of model accuracy of FGSM with the change of attack parameters. When the raw accuracy is stable at 98.33%, when the perturbation strength (ε) is at a low level, i.e., increasing from 0 to 0.01, the countermeasure accuracy plummets to 42.00%, and the success rate of the attack rises sharply to 57.29%, which fully demonstrates the powerful destructive nature of the FGSM attack under low perturbation strength, and the very small perturbation strength of 0.001 makes the model accuracy drop significantly. As the perturbation strength is further increased from 0.05 to 0.3, the confrontation accuracy rate stabilises at 39.00% after dropping to 36.67%, and the success rate of the attack likewise stabilises at 60.34%, which shows a decreasing marginal effect under high perturbation strength, strongly proving the high efficiency of the FGSM attack. These data also clearly show that the traditional training approach does not take into account the adversarial perturbation, resulting in the model showing significant vulnerability in the face of adversarial attacks.

3.3 Defense research

Table 2. Prediction accuracy of different defence methods for generating adversarial samples

Mode of defence	Accuracy

None	0%
Data preprocessing	9.00%
PGD Adversarial Learning (100 random adversarial samples))	18.00%
PGD Adversarial Learning (500 random adversarial samples)	26.00%
PGD Adversarial Learning (1000 random adversarial samples)	23.00%

Table 2 shows that the model accuracy is 0% when there is no defence, and the data preprocessing only improves it to 9.00%, which shows that the defence effect is poor. PGD Adversarial Learning performs outstandingly, with an accuracy of 18.00% using 100 random adversarial samples, and it rises to 26.00% when it is increased to 500, which proves that increasing the number of samples can effectively improve the defence accuracy, and it is a feasible defence strategy. However, when the samples are increased to 1000, the accuracy rate drops to 23.00%. This shows that PGD adversarial learning is a feasible defence method to defend against attacks, and its accuracy rate is higher than that of general processing methods but there are some bottlenecks.

The above data clearly demonstrates the powerful effectiveness of the attack in breaking through the defence system, and shows the performance of the adversarial learning method in defending against the attack, which provides a new perspective for understanding the nature of attack and defence in depth. However, there are still some limitations in this study. On the one hand, the size and diversity of the LFW dataset are insufficient, which limits the generalisation ability of the model in a wider range of scenarios. On the other hand, the PGD and FGSM algorithms have their own limitations, with high computational complexity in the PGD algorithm and a single direction of perturbation in the FGSM attack, which affects the comprehensiveness of attack and defence research. In addition, the research on adversarial attack and defence needs to make breakthroughs at multiple levels: on the one hand, the universality of adversarial samples in complex tasks and cross-modal scenarios should be explored in depth, and assessment benchmarks closer to the complex conditions of the physical world should be constructed; on the other hand, it is necessary to break through the limitations of the current defence strategies, and to improve the robustness through theoretical analyses and dynamic self-adaptive mechanisms, while strengthening cross-disciplinary collaboration, and pushing the unifying theoretical framework to be establishment[3].

4 Conclusion

In this paper, the VGG model of convolutional neural network and the counterattack method used to study the face recognition model in terms of attack and defence against attack. Through the study data, it is found that in the case of the original accuracy remains unchanged at 98.33%, the FGSM algorithm, with the increase of perturbation parameter, the model correctness rate shows a cliff decline to 42.00% first, and then a slow decline until a smooth 39.00%, which proves the high efficiency of the FGSM attack and the model

shows significant vulnerability in the face of the antagonistic attack. Meanwhile, the ordinary defence means of data preprocessing has a very poor defence effect with only 9.00% accuracy rate, while the PGD adversarial learning defence method, with the increase of the number of adversarial samples, improves the accuracy rate to a maximum of 26.00%, which proves that the adversarial learning defence method has a good defence against the attack, and at the same time, there are certain bottlenecks in defenc.

This study systematically reveals the significant vulnerability of well-trained face recognition models under adversarial attack scenarios, and at the same time verifies the existence of protection efficacy boundaries for current mainstream adversarial learning defence schemes. This finding provides an important warning for industry practitioners and has guiding value for improving the practical security of biometric authentication systems. Future research should build multi-scenario cross-modal face evaluation benchmarks, integrate physical noise to improve test effectiveness; optimize the effectiveness of attack algorithms, enhance adversarial performance through semantic perturbation and distributed computing, and explore the biological mechanism of model vulnerability; establish attack and defence dynamic game models with provable robust boundaries, and develop theory-driven defence strategies; promote the synergistic innovation of cryptography, hardware security, and defence technology, and build an open attack and defence platform, and realize the system upgrade of face recognition system from passive defence to active defence.

References

1. Gu, J. et al., "Recent advances in convolutional neural networks," Pattern Recognition, vol. 77, pp. 354-377, 2018, doi: 10.1016/j.patcog.2017.10.013.
2. Wang, R. et al., "Amora: Black-box Adversarial Morphing Attack," in Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 2020. [Online]. Available: https://doi.org/10.1145/3394171.3413544.
3. Akhtar, N. and Mian, A., "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," IEEE Access, vol. 6, pp. 14410-14430, 2018, doi: 10.1109/ACCESS.2018.2807385.
4. Goodfellow, I. J., Shlens, J. and Szegedy, C., "Explaining and Harnessing Adversarial Examples," arXiv preprint, arXiv:1412.6572, 2014, doi: 10.48550/arXiv.1412.6572.
5. Paricio-Garcia, A., Lopez-Carmona, M. A., Sierra-Arquero, S. and Manglano-Redondo, P., "Analysis and evaluation of autoencoder-driven dimensionality reduction for face recognition pipelines," Applied Soft Computing, vol. 172, p. 112877, 2025, doi: 10.1016/j.asoc.2025.112877.
6. Gewers, F. L. et al., "Principal Component Analysis: A Natural Approach to Data Exploration," ACM Computing Surveys, vol. 54, no. 4, Article 70, 2021, doi: 10.1145/3447755.
7. Wang, W., Shen, J. and Ling, H., "A Deep Network Solution for Attention and Aesthetics Aware Photo Cropping," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 7, pp. 1531-1544, 2019, doi: 10.1109/TPAMI.2018.2840724.
8. Yoshida, M., Namura, H. and Okuda, M., "Adversarial Examples for Image Cropping: Gradient-Based and Bayesian-Optimized Approaches for Effective Adversarial Attack," IEEE Access, vol. 12, pp. 86541-86552, 2024, doi: 10.1109/ACCESS.2024.3415356.

9. Zhang, K., Zuo, W., Chen, Y., Meng, D. and Zhang, L., "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3142-3155, 2017, doi: 10.1109/TIP.2017.2662206.
10. Ding, C. and Tao, D., "Robust Face Recognition via Multimodal Deep Face Representation," IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2049-2058, 2015, doi: 10.1109/TMM.2015.2477042.
11. Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint, arXiv:1409.1556, 2014, doi: 10.48550/arXiv.1409.1556.
12. Hussain, M., Thaher, T., Almourad, M. B. and Mafarja, M., "Optimizing VGG16 Deep Learning Model with Enhanced Hunger Games Search for Logo Classification," Scientific Reports, vol. 14, no. 1, p. 31759, 2024, doi: 10.1038/s41598-024-82022-5.
13. Waghela, H., Sen, J. and Rakshit, S., "Robust Image Classification: Defensive Strategies Against FGSM and PGD Adversarial Attacks," arXiv preprint, arXiv:2408.13274, 2024, doi: 10.48550/arXiv.2408.13274.
14. Takahashi, R., Matsubara, T. and Uehara, K., "Data Augmentation Using Random Image Cropping and Patching for Deep CNNs," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 9, pp. 2917-2931, 2020, doi: 10.1109/TCSVT.2019.2935128.
15. Machado, G. R., Silva, E. and Goldschmidt, R. R., "Adversarial Machine Learning in Image Classification: A Survey Toward the Defender’s Perspective," ACM Computing Surveys, vol. 55, no. 1, Article 8, 2021, doi: 10.1145/3485133.