Research on The Security of Face Recognition Systems Based on Digital and Physical Counterattacks

Tengyan Ma

Ulster College, Shaanxi University of Science & Technology, Xi'an, Shaanxi, 710021, China

Abstract. Adversarial attacks have become an important direction in the research of their security by generating adversarial samples or physically interfering to deceive face recognition systems. This study compares the two methods of digital and physical attacks, aiming to evaluate their effects and differences in practical applications. In this paper, the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) methods are used to generate antagonistic samples, and physical attacks are simulated by adding eyeglasses stickers in the digital environment. The experimental results show that the FGSM and PGD attacks reduce the model accuracy from 97.70% to 42.45% and 21.58%, respectively, while the physical attack causes the accuracy to drop to 71.25% by adding eyeglasses stickers, which verifies that the adversarial attack is a significant threat to the face recognition system. These findings explain the potential threat of adversarial attacks on face recognition systems and can provide an important basis for improving system security.

1 Introduction

Face recognition technology, as a core breakthrough in the field of biometrics, has been widely used in areas such as security monitoring, financial payment and intelligent terminals. However, with the large-scale deployment of deep neural networks, its security is facing serious challenges. In particular, adversarial attacks, which deceive the face recognition system by generating adversarial samples or physical interference, have become a research hotspot in this field [1]. Such attacks not only seriously threaten the security of personal privacy, but also may trigger risks such as the failure of identity authentication systems in public places. For example, attackers can successfully break through the face verification system of mobile payment through digital adversarial samples [2], or wear specially designed adversarial glasses to deceive the smart security check channel in airports [3], and these security incidents have caused direct economic losses and triggered a public trust crisis. Therefore, in-depth study of the mechanism of adversarial attacks and a comparison of the differences between the two attacks have become an important issue to ensure the trustworthy development of AI.

202315030403@sust.edu.cn

In recent years, face recognition models based on deep features have progressed significantly. The advanced model represented by ArcFace achieves 99.83% accuracy on the Labelled Faces in the Wild (LFW) dataset by introducing an additive angular interval loss function [4]. However, this deep learning based recognition system still faces serious challenges against attacks. Digital and physical attacks are the two main ways to counter attacks. Digital attacks manipulate the input data through pixel-level perturbations, and typical methods include the Fast Gradient Sign Method (FGSM) algorithm based on single-step gradient ascent and the Projected Gradient Descent (PGD) algorithm using iterative optimisation [5, 6]. Physical attacks, on the other hand, alter facial features by adding interference to the face image, such as the 'antagonistic glasses' designed by Sharif et al. that successfully mislead the recognition system in real scenes by interfering with special patterns [7]. Although both attacks can successfully deceive the face recognition system, there are significant differences between them in terms of attack implementation, applicable scenarios and defence difficulty.

This paper aims to reveal the core differences between digital and physical attacks through quantitative experiments. Section 2 of the paper describes the experimental design method and evaluation index in detail; Section 3 presents the comparative experimental results of digital and physical attacks; Section 4 summarises the research conclusions and future directions. This research is expected to provide a theoretical basis for constructing a dynamic defense system and promote the evolution of face recognition technology toward a more secure and reliable direction.

2 Data and methods

2.1 Dataset and selected face recognition system

This paper is based on the analysis of the LFW dataset from http://vis-www.cs.umass.edu/lfw/, which contains 13,233 face images covering 5,749 different people, and its distinctive features are the high diversity of lighting, pose, occlusion, and expression, and more than 83% of the identities have sample sizes of only 1-5 images, which can effectively simulate the real scene recognition challenges [8]. The face recognition system is based on the pre-trained InceptionResnetV1 model from the VGGFace2 dataset, which fuses the multi-scale convolution of the Inception module with the residual concatenation of ResNet, balancing the computational efficiency and feature expressiveness through the 160*160 pixel input size while maintaining the 152-layer depth structure [9].

The data were pre-processed before the experiment to eliminate LFW image resolution differences by forcing the image size to be uniform to 160*160, and the ToTensor transformation was used to normalise the pixel values to the range of [0,1], which was strictly aligned with the data distribution in the pre-training phase. The preprocessing process did not introduce additional normalisation operations to maintain consistency with the original VGGFace2 training protocol. This processing aims to ensure that the input tensor dimensions are fully adapted to the model architecture while eliminating feature extraction bias due to differences in the original data sizes or value distributions.

2.2 Methods of attack

In this paper, two methods, digital and physical attacks, are used to systematically evaluate the accuracy of face recognition models under input perturbation by generating adversarial samples and physical interference means. The digital attack is based on the gradient

optimisation method implementing FGSM and PGD attacks, where FGSM generates adversarial samples through single-step gradient sign perturbation, while PGD uses an iterative gradient ascent strategy combined with ε-ball neighborhood constraints for adversarial Optimisation.

In physical attacks, interference in a realistic scene is simulated by adding eyeglass stickers to the original character image in a digital environment. The method effectively changes the model discrimination by adjusting the sticker size and optimising the transparency so that the image maintains its natural appearance while at the same time. The experimental flow is shown in Fig. 1.

Fig.1 Experimental procedure(Photocredit : Original)

3 Analysis of results

3.1 Analysis of attack results

Table 1 Accuracy and success rate before and after the attack

Type of attack	Post-attack accuracy	Attack success rate
No attack	97.7%	—
FGSM（ε=0.03）	42.45%	57.55%
PGD（ε=0.03）	21.58%	78.42%
Physical attack	71.25%	—

In this study, the accuracy of face recognition system recognition was tested by digital attacks (FGSM attack and PGD attack) and physical attack by adding eyeglasses stickers. As shown in Table 1, the original model has an accuracy of 97.70% on the LFW dataset. In the digital attack, the model accuracy drops to 42.45% and the attack success rate is 57.55% for the FGSM attack, while the PGD attack achieves an even higher attack success rate of 78.42% through an iterative optimisation strategy, and the model accuracy further drops to 21.58%. This result suggests that PGD can generate more effective adversarial samples due to the nature of multi-step gradient optimisation, but its computational cost is significantly higher than that of the single-step FGSM. In terms of physical attacks, the accuracy of the model decreases from 97.70% to 71.25% with the addition of the eyeglasses sticker. Thus, the impact of physical attacks on the model recognition ability is relatively weak compared to digital attacks.

Fig. 2 demonstrates the visual comparison between the original image and the antagonistic sample, where it can be seen that the sticker covers the bridge of the nose and the area around the eyes, resulting in pixel shifts at key facial feature points. At the same time, the high contrast stripes of the sticker create a chromatic conflict with the background skin colour, interfering with the model's perception of skin colour continuity. Although

physical attacks are less disruptive than digital attacks, the fact that they do not require direct access to model inputs makes them more threatening in realistic scenarios.

Fig.2 Comparison between the original images and the adversarial examples(Photocredit : Original)

Fig. 3, Fig. 4 and Fig. 5 further compares the noise distributions of the different attacks: the FGSM perturbation set is finally localised to a region, the PGD perturbation is more homogeneous and stronger, whereas the noise of the physical attack appears as a discrete region covered by stickers.

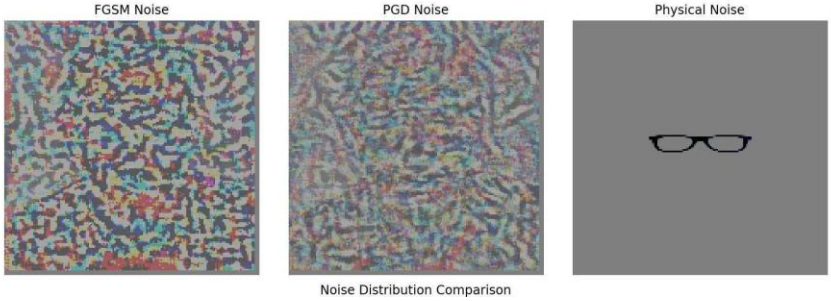

Fig.3 Comparison of Visualizations and Distributions of FGSM, PGD, and Physical Noise(Photocredit : Original

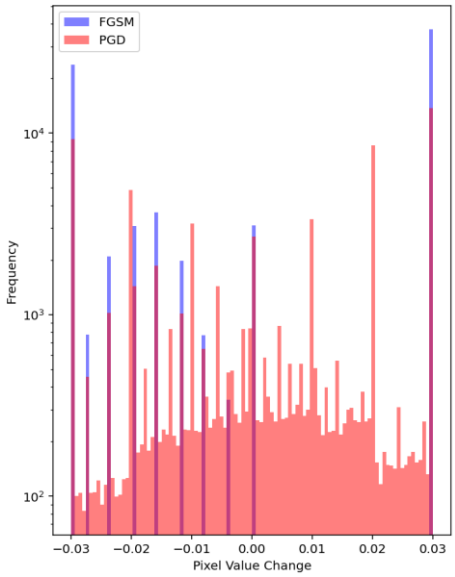

Fig.4 Histogram of digital attack noise distribution(Photocredit : Original)

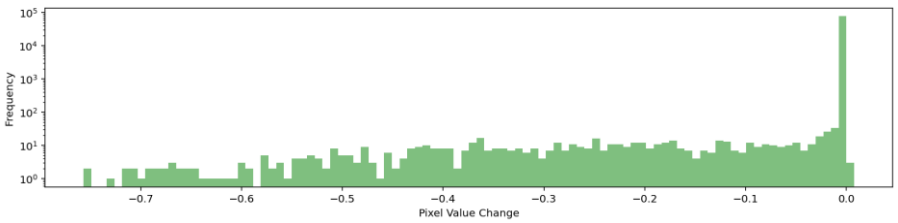

Fig.5 Histogram of physical attack noise distribution(Photocredit : Original)

3.2 Discussion

This study reveals the different threatening nature of digital and physical attacks in face recognition systems. Digital attacks (e.g., FGSM, PGD) generate antagonistic samples by directly manipulating the pixel values of the input data, and their efficiency stems from the precise use of model gradients, but their limitations lie in their reliance on white-box assumptions and the ease with which the generated perturbations can be eliminated by standardised preprocessing. In contrast, physical attacks (e.g., adding eyeglasses stickers) have a lower success rate, but they are more difficult to detect through physical media and have stronger practical value in real security scenarios. The innovation of this study lies in the construction of an agreement assessment framework for digital and physical models and quantitatively comparing the noise distribution characteristics of the two.

However, this paper still suffers from several limitations. The experiments only use static eyeglass stickers as physical attack vectors and do not consider the effect of dynamic interference on the system. Moreover, the LFW dataset does not contain samples of novel occlusions despite its scene diversity. Future research can be conducted by combining Generative Adversarial Network (GAN) techniques to generate more occluded adversarial samples, or using reinforcement learning to dynamically adjust the physical attack

parameters in order to improve the adaptability and covertness of the attack. Attack scenarios can also be extended to explore multimodal attack tactics, such as combining voice or action jamming, and investigating the effect of light variations on the robustness of physical attacks [10-13].

4 Conclusion

In this study, the threat of adversarial attacks on face recognition systems is verified through digital and physical attacks. Based on the LFW dataset and the InceptionResnetV1 model, the accuracy and noise distribution characteristics of the model before and after the attack are tested using a unified evaluation framework, and the mechanism of different attack modes is revealed by combining quantitative and qualitative analyses. The results show that the adversarial attacks are significantly destructive to the face recognition system. In the digital attack, both FGSM and PGD attacks can effectively generate adversarial samples and significantly reduce the accuracy of the model, the FGSM attack makes the accuracy drop to 42.45%, while the PGD attack outperforms the FGSM due to the iterative optimisation feature and further depresses the accuracy to 21. 58%, but with higher computational cost. The physical attack destroys the recognition ability of the model by adding eyeglass stickers, which decreases the accuracy to 71.25%. Although the decrease (26.45%) is lower than the digital attack, its nature of directly interfering with the physical imaging without invading the model's interior makes it more feasible to implement in real-world scenarios. These findings provide an important basis for security improvements in face recognition systems, and also highlight the need to introduce robustness in the model design and training process.

References

1. Akhtar, N., Mian, A.: "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," arXiv.org, 2018.
2. Xu, Y., Raja, K., Ramachandra, R., Busch, C.: "Adversarial Attacks on Face Recognition Systems," in Advances in Computer Vision and Pattern Recognition, 2022, pp. 139 – 161.
3. Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., Chen, P., Wang, Y., Lin, X.: "Adversarial T-Shirt! Evading Person Detectors in a Physical World," in Lecture Notes in Computer Science, 2020, pp. 665 – 681.
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690 – 4699.
5. Goodfellow, I. J., Shlens, J., Szegedy, C.: "Explaining and Harnessing Adversarial Examples," arXiv.org, 2014.
6. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: "Towards Deep Learning Models Resistant to Adversarial Attacks," arXiv.org, 2017.
7. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M. K.: "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16), Association for Computing Machinery, New York, NY, USA, 2016, pp. 1528 – 1540.

8. Huang, G. B., Mattar, M., Berg, T., Learned-Miller, E.: "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," 2008.
9. Cao, Q., Shen, L., Xie, W., Parkhi, O. M., Zisserman, A.: "VGGFace2: A Dataset for Recognising Faces Across Pose and Age," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 67-74, IEEE.
10. Song, Y., Shu, R., Kushman, N., Ermon, S.: "Constructing Unrestricted Adversarial Examples with Generative Models," arXiv.org, 2018.
11. Kim, T., Yu, Y., Ro, Y. M.: "Defending Physical Adversarial Attack on Object Detection via Adversarial Patch-Feature Energy," in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1905 – 1913.
12. Tabassi, E.: "A Taxonomy and Terminology of Adversarial Machine Learning," 2023.
13. Vakhshiteh, F., Nickabadi, A., Ramachandra, R.: "Adversarial Attacks Against Face Recognition: A Comprehensive Study," IEEE Access, 2021, 9, pp. 92735 – 92756.