RESEARCH ON WHITE-BOX COUNTER-ATTACK METHOD BASED ON CONVOLUTION NEURAL NETWORK FACE RECOGNITION

Shuya Tian and Xiangwei Lai

College of Computer and Information Sciencce, Southwest University, Chongqing, China

ABSTRACT

In recent years, deep neural network has been widely used in face recognition, in which the model of a convolution neural network for face recognition is mostly black box model. Because the model structure and related parameters cannot be obtained, the attack effect of the counter sample is poor. In order to better realize the attack effect of the black box attack, this paper uses the white box attack to realize the black box attack. Aiming at the convolution neural network face recognition model, this paper proposes an improved FGMS counterattack algorithm, which uses the cosine similarity between the clean sample and the antagonistic sample as the loss function. The threshold is set to 0.8 as the condition for the success of the attack. In order to avoid excessive changes in the image, the threshold super-parameters is set to limit the range and size of the disturbance fluctuation, so that the countermeasure samples are not easy to be detected and improve the visual quality. Countermeasure samples are detected by black box attack on the VGG16 model, and a good attack effect is obtained.

KEYWORD

Face recognition; adversarial examples; White box attack

1. INTRODUCTION

In recent years, face recognition model based on depth neural network is widely used in various scenes. However, with the deepening of the research, the researchers found that through the imperceptible modification of a small number of pixels, generated confrontation samples will lead to the misclassification of the image of the deep learning recognition model. Take the face recognition system as an example, because its main function is for identity authentication, it is likely to lead to loopholes being exploited due to the neglect of its weak defense, resulting in personal privacy information disclosure, threats to property security and other security problems. Therefore, security of face recognition model is a research field that needs great attention.

Most of the actual face recognition models are black-box models, that is, we can not directly obtain the network structure, data sets, model parameters and other information of the model. There are many algorithms for black-box attacks, one of which is to realize black-box attacks with the help of white-box attacks [1], that is, to find a model that has the same task as the target model to be attacked. at the same time, the model with the details of the model is used as an alternative model, and then the white-box attack algorithm is used to replace the model to generate confrontation samples, and finally the confrontation samples are used as the input of the target model to realize the black-box attack. Therefore, in order to obtain countermeasure samples with better attack effect, it is necessary to study the white-box attack face recognition model. Convolution Neural Network (Convolutional Neural Networks, CNN) is the most typical and universal deep neural network algorithm, which is used in many fields. Therefore, this paper

David C. Wyld et al. (Eds): NLDML, CCITT-2022
pp. 43-51, 2022. IJCI – 2022 DOI:10.5121/ijci.2022.110204

mainly focuses on the white-box attack algorithm of face recognition model based on convolution neural network.

2. RELATED RESEARCH BACKGROUND

The concept of "antagonistic sample" was first put forward by SzegedyC et al. [2], New samples are obtained by adding small disturbances to the data set, which makes the output of the deep learning model misclassified, and the anti-sample attacks can be divided into target attacks and non-target attacks. In addition, according to the attack environment, it can also be divided into white-box attack and black-box attack [3]. At present, there have been many achievements in the research of anti-attack methods for face recognition.

In the black box attack, the attacker cannot get the key details of the attack model and can only reason some characteristics of the model according to the input and output information [4]. In contrast, in a white-box attack, the attacker knows the details of the attack model, such as data preprocessing methods, model structure, model parameters, and in some cases, the attacker can master some or all of the training data information [4]. Szegeedy [5] proposed the L-BFGS method, which makes the neural network model misclassified by adding a slight disturbance to the sample. Goodfellow et al proposed FGSM [6], namely fast gradient iterative method (Fast Gradient Sign Method), which produces the disturbance of anti-attack effect in the gradient direction, and modifies the original image to form countermeasure samples.

Kur-akin [7] proposed the BIM (Basic Iterative Methods) & ILLC (Iterative Least Likely Class) algorithm, which adds multiple input parameters in the gradient direction and uses a small search step to iteratively calculate the disturbance to obtain more aggressive countermeasure samples. Moosavi-Dezfooli et al. [8] proposed the DeepFool method, which uses the idea of hyperplane classification and uses less disturbance to achieve an effect similar to that of FGSM attacks. Carlini et al. [7] proposed a cymbal W attack, which makes the generated countermeasure samples difficult to detect by limiting the norm, and achieves a strong attack effect.

For it is difficult to obtain the internal structure information of the attack model, it is difficult for attackers to design aggressive confrontation samples by using black-box attacks, while the method based on white-box attacks can only attack known models with poor generalization ability and poor migration. Unable to attack the model in the real scene. Therefore, the main research focus of this paper is to combine the idea of white-box attack and black-box attack, and propose a white-box attack on the alternative model based on the improved fast gradient symbol (FGSM) algorithm, and use the confrontation samples to attack the target model to realize the black-box attack.

3. FACE RECOGNITION MODEL BASED ON CONVOLUTION NEURAL NETWORK

The fragility of the face model based on convolution neural network is studied, and the attack effect of black box attack is verified by white box attack. In this paper, the VGG16 model is designed as the target model to be attacked, and the convolution neural network face recognition model which is consistent with the target model is built as an alternative model. Among them, VGG16 is a classical convolution neural network model, and 16 represents the depth of the model, which is superimposed by 13 convolution layers and 3 fully connected layers[9].

3.1. MTCNN（Multi-Task Cascade Convolutional Networks）

MTCNN [10] algorithm can perform face detection and face alignment at the same time. The MTCNN algorithm consists of three cascaded networks: P-Net (Proposal Network), R-Net (Refine Network) and O-Net (Output Network).

Among them, the main purpose of P-Net network is to obtain the candidate box and feature extraction of the human face. During the training, the face classifier is used to determine whether there is a face in this image, and the frame regression and the location of face key points are given. R-Net is a further optimization of P-Net to refine the input and delete a large number of non-human face frames. O-Net has another convolution layer than R-Net, which has more supervision on the face region, more accurate processing results, and more output of key points.

3.2. Convolution Neural Network Face Recognition Model

In this paper, the convolution kernel with the size of 3X3 is used in the model training, and the depth is increased by superimposing the convolution layer. The specific setting of the hidden layer is shown in Table 1.

Table1. The detail parameters of based on face recognition of convolution neural network

Name	Type	Input size	Output size	Convolution kernel
Conv2d_1	Convolution	150*150*3	150*150*32	32*3*3
Conv2d_2	Convolution	150*150*32	150*150*32	32*3*3
Max_poolong2d_1	MaxPooling	150*150*32	75*75*32	-
Conv2d_3	Convolution	75*75*32	75*75*64	64*3*3
Conv2d_4	Convolution	75*75*64	75*75*64	64*3*3
Max_poolong2d_2	MaxPooling	75*75*64	37*37*64	-
Conv2d_5	Convolution	37*37*64	37*37*128	128*3*3
Conv2d_6	Convolution	37*37*128	37*37*128	128*3*3
Max_poolong2d_3	MaxPooling	37*37*128	18*18*128	-
Flatten_1	Flatten	18*18*128	41472	-
Dense_1	Dense	41472	256	-
Dropout_1	Dropout	256	256	-
Dense_2	Dense	256	18	

The purpose of introducing the pool layer is to reduce the size of the feature graph produced by the convolution layer and reduce the dimension of the convolution results. The convolution kernel parameters in the hidden layer are shared and the connection between each layer is sparse, which can reduce the amount of computation and improve the operation speed. At the same time, it has a good learning effect for image features. Usually, a pooling layer is introduced between convolution layers in order to reduce some parameters and prevent overfitting to some extent. The fully connected layer is usually placed in the last part of the hidden layer to map the learned feature representation of the tag space of the face sample. The whole face recognition process is shown in Figure 1.

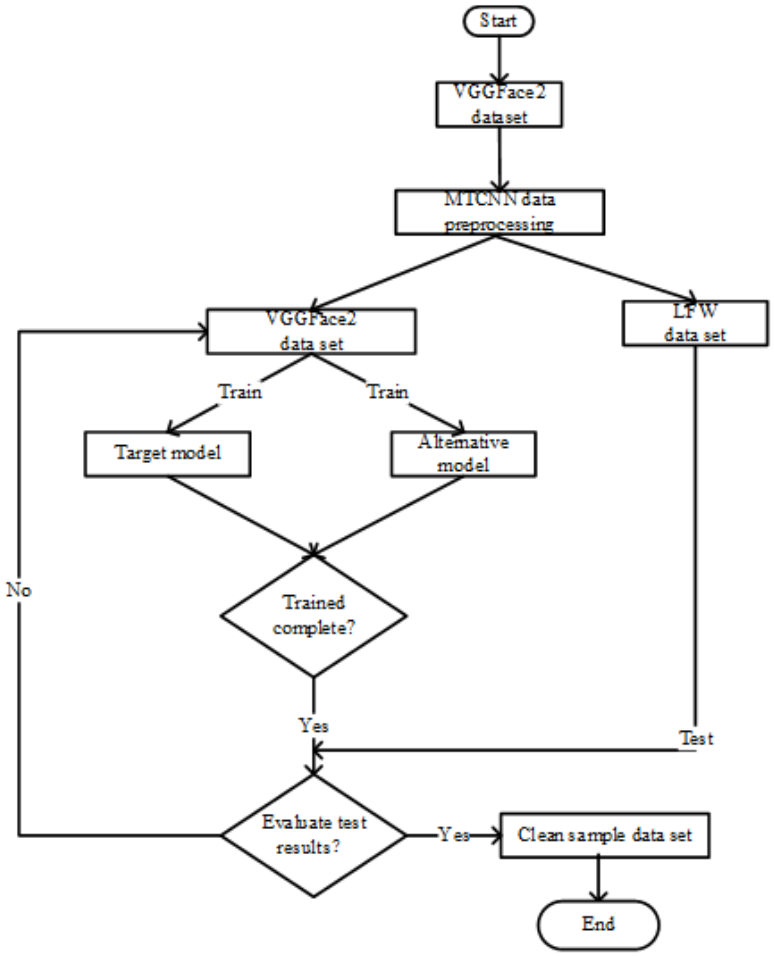

Figure 1. flow chart of face recognition model

The recognition effect of the LFW verification set on the alternative model is shown in Figure 2. Among them, the ordinate represents the recognition rate of the LFW verification set on the lifting point model, and the abscissa represents the epoch value of the alternative model. As can be seen from the chart, the highest recognition rate of the LFW verification set is 94.45%, indicating that the model itself has a good recognition effect and can be used normally in subsequent anti-attack experiments.

4. ANTI-ATTACK ALGORITHM BASED ON IMPROVED FGMS

The method based on gradient attack is to add disturbance and increase noise to the global feature space of the original sample along the gradient direction, which is easy to cause some noises that are too obvious and easy to be detected by people. At the same time, considering that the network of convolution neural network model for face recognition is more complex. In this paper, an anti-attack algorithm based on improved FGMS is proposed. That is, the idea of FGSM algorithm is further optimized, the cosine distance which is more suitable for face comparison is taken as the countermeasure loss, and the super marketer α is introduced to limit the maximum intensity of the disturbance to ensure that the disturbance of each pixel of the original sample does not exceed α. In order to obtain an imperceptible confrontation sample. At the same time, the gradient attack method aims at the global feature space of the original sample. In order to avoid the interference of background and other objects, we process the data set through MTCNN, and finally get the

image of only the face part of the size of 112X112. The attack process is shown in Figure 3. The generated confrontation sample is shown in Figure 4.

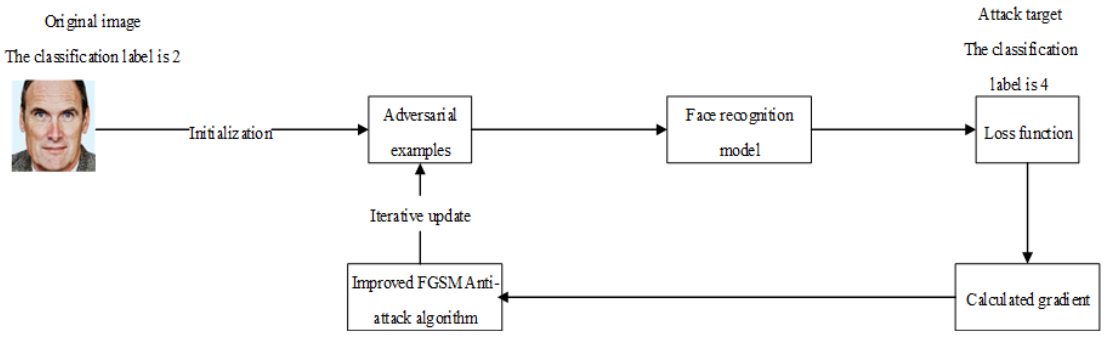

Figure 3. flowchart of counterattack

Figure 4. generated confrontation sample

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Experimental Setup

Experimental environment: The framework used in the experiment are tensorflow and keras, and the language used in the experiment is python, and the library used of python include Numpy, skimage, opencv, PIL, matplotlib and so on.

Data set：

1) VGGface2 [11] selects 18 categories, each of which contains an average of 300 pictures as training data, and randomly divides the training data set into training set and verification set according to the proportion of 9:1. Among them, the training set contains 5329 images, and the verification set contains 583 images. Finally, the training parameters are saved.
2) LFW[12] randomly selects the images corresponding to 18 people in the training data set (each identity corresponds to at least two face images) as the test set to be recognized, and ensures that the model has not learned these clean samples to be attacked, which is used to test the recognition accuracy of the face recognition model in this paper.

The face images of the two data sets are preprocessed by MTCNN, which is scaled to the size of 112X112 and contains only the part of the face image, so as to reduce the interference caused by the background and other objects in the face recognition model training. As shown in Figure 5. All the face images in the test set to be recognized can be correctly identified by using the alternative model and the target model.

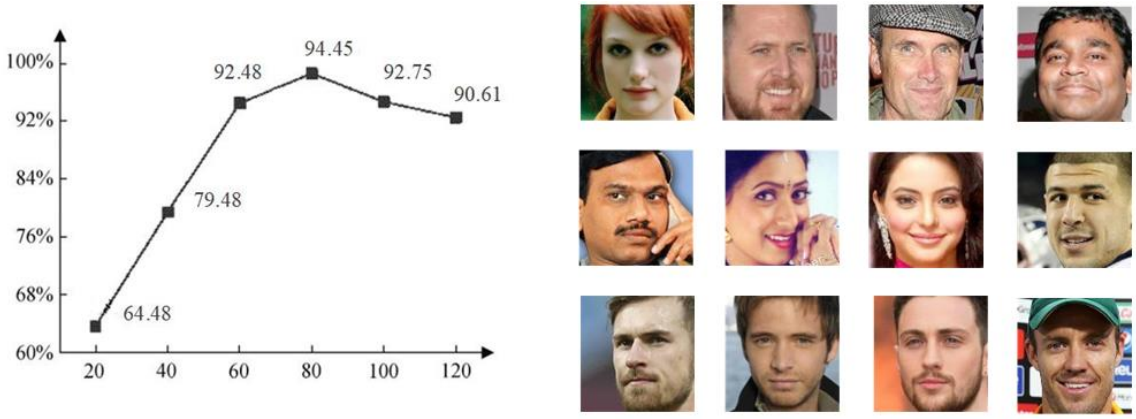

Figure 2 recognition rates of LFW verification set Figure 5. VGGFace2 face images aligned with MTCNN

5.2. Evaluation Index

In this paper, the cosine similarity is used to judge the identity information. Therefore, baseline similarity (BS) [13], final similarity（FS）, structural similarity index between the original image and antagonistic sample (SSIM）[14], and the success rate of attack as the evaluation index.

Among them, the baseline similarity refers to the cosine similarity between the clean image and the attack target, such as (1). The final similarity refers to the cosine similarity between the countermeasure sample and the attack target. The greater the final similarity, the greater the probability that the confrontation sample and the attack target will be identified as the same person, such as (2). The result similarity refers to the cosine similarity between the clean image and the antagonistic sample. In this paper, the threshold is set to 0.8 as a condition to judge the success of the attack. That is, if the result similarity is less than 0.8 or the identity corresponding to the current result similarity is consistent with the identity of the attack target, the attack is judged to be successful, such as (3):

$$BS = \frac{\sum_{i=1}^{n}(e_{ai} * e_{li})}{\sqrt{\sum_{i=1}^{n}((e_{ai})^2)} * \sqrt{\sum_{i=1}^{n}((e_{li})^2)}} \tag{1}$$

$$FS = \frac{\sum_{i=1}^{n}(e_{ami} * e_{li})}{\sqrt{\sum_{i=1}^{n}((e_{ami})^2)} * \sqrt{\sum_{i=1}^{n}((e_{li})^2)}} \tag{2}$$

$$FS = \frac{\sum_{i=1}^{n}(e_{ami} * e_{ai})}{\sqrt{\sum_{i=1}^{n}((e_{ami})^2)} * \sqrt{\sum_{i=1}^{n}((e_{ai})^2)}} \tag{3}$$

e_{ai} represents the I-dimensional vector value of the feature vector extracted by the model of the original image. e_{li} represents the I-dimensional vector value of the feature vector extracted by the model of the attack target. e_{ami} represents the value of the I-dimensional vector of the feature vector extracted by the countermeasure sample through the model.

The structural similarity is based on three distance measures: brightness $l(x,y)$, contrast $c(x,y)$ and structure $s(x,y)$ of the original image x and the antagonistic sample y, such as (4).

$$SSIM(x,y) = l(x,y)c(x,y)s(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (4)$$

μ_x is the mean of x, μ_y is the mean of y, σ_x^2 is the variance of x, σ_y^2 is the variance of y, σ_{xy}^2 is the covariance of xy, $c_1 = (k_1 L)^2, c_2 = (k_2 L)^2, c_3 = c_2/2$, L is the range of pixel values, L=1, k_1=0.01, k_2=0.03。From this, we can get $SSIM(x,y) \in [0,1]$。

The success rate of attack refers to the number of successful confrontation samples divided by all confrontation samples.

5.3. Analysis of the Experimental Process and Results

In this paper, the VGG16 model is taken as the target model to be attacked, and the convolution neural network face recognition model which is consistent with the function of the target model is built as an alternative model. The counterattack algorithm based on the improved FGSM attacks the alternative model to generate antagonistic samples, and the antagonistic samples are attacked by black box on the target model. The specific process is shown in Figure 6.

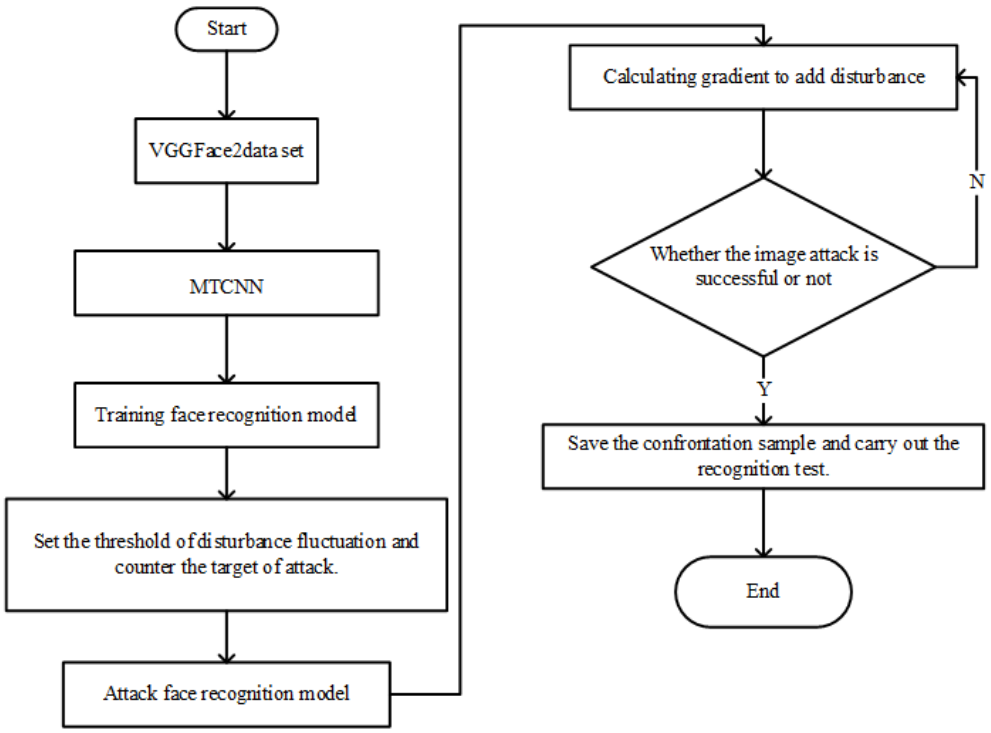

Figure 6. overall flow chart

In the training process of the face recognition model, the batch size is set to 32, and the data is recorded in each training round. And use the recognition accuracy of the verification set

calculated by each epoch to determine whether it is necessary to stop training in advance, so as to prevent the over-fitting phenomenon.

After the model training is completed, the LFW verification set is used to test the model recognition effect. The test method is to judge whether the face recognition results of 18 groups in LFW are correct, and finally calculate the accuracy. Both the attack model and the VGG16 model have a recognition rate of more than 94% on the LFW verification set, as shown in Table 2, which shows that the model itself has a good recognition effect and can be used normally in subsequent anti-attack experiments.

Table 2. recognition rate of human face recognition model on LFW data set

	Alternative model	target model
LFW	94.45%	95.30%

In the test set, one identity is randomly selected as the attacker, and the remaining identity as the attack target. The sign of the success of the attack is that the similarity of the result is less than 0.8, or the identity corresponding to the similarity of the current result is consistent with the identity of the attack target. During each attack, the baseline similarity remains the same. Ten identities are randomly selected from the LFW verification set as attackers to experiment. Table 3 describes the average success rate and average SSIM of LFW verification sets in alternative models and model recognition rates and when both models are attacked. The experimental results show that the countermeasure samples generated by the attack substitution model based on the improved FGMS counterattack algorithm successfully attack the target model. At the same time, compared with the target attack, the effect of non-target attack is more significant.

Table 3. performance of human face recognition model

	LFW	Target attack	Non-target attack	Average SSIM
Alternative model	96.78%	86.78%	96.01%	0.77
target model	98.30%	84.30%	93.30%	0.71

6. CONCLUDING REMARKS

The experimental results show that the countermeasure samples based on the improved FGMS attack algorithm can achieve better attack results on the complex VGG16 model, but the cosine similarity between the attacker and the attack target is small, so it is difficult to get the corresponding countermeasure samples to achieve the target attack. At the same time, the number of iterative rounds of the attack will also have an impact on the effectiveness of the attack. Directly add disturbance to the image, a wide range of search, time-consuming, some cases need to add a large number of disturbances to achieve the attack, which will cause visual quality discomfort. The introduction of the super-parameter α limit the threshold of disturbance fluctuation reduces the problem of visual quality discomfort to a certain extent, but at the same time increases the consumption of time cost, and has a certain impact on the formation of target antagonistic samples. The next step is to introduce attention mechanism and other methods to limit the scope and location of disturbance, so as to further solve the problem of visual quality discomfort. At the same time, the confrontation samples are added to the training data set for confrontation training, in order to resist other confrontation samples and enhance the robustness and data security of the face recognition model.

ACKNOWLEDGEMENTS

This work was supported by program of “The Construction of Big Data Platform for Early Childhood Education and its Application in Promoting the Balanced Development of Regional Education", sponsored by the "Program of Technology Innovation and Application Development [The Special Program for Alleviating Poverty by Technology])" by Chongqing Committee of Science and Technology (Program No. CSTC2019JSCX-KJFP0004)

REFERENCES

[1] Dong, Y., et al., Efficient Decision-based Black-box Adversarial Attacks on Face Recognition. 2019.
[2] Szegedy, C., et al., Intriguing properties of neural networks. 2013.
[3] Papernot, N., et al., The Limitations of Deep Learning in Adversarial Settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 2016: p. 372-387.
[4] Ryu, G., H. Park and D. Choi, Adversarial attacks by attaching noise markers on the face against deep face recognition. Journal of information security and applications, 2021. 60: p. 102874.
[5] Szegedy, C., et al., Intriguing properties of neural networks. 2014.
[6] Goodfellow, I., J. Shlens and C. Szegedy, Explaining and Harnessing Adversarial Examples. arXiv 1412.6572, 2014: p. {}.
[7] Kurakin, A., I. Goodfellow and S. Bengio, Adversarial Examples in the Physical World. 2018. p. 99-112.
[8] Moosavi-Dezfooli, S., A. Fawzi and P. Frossard, DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. 2016. p. 2574-2582.
[9] Liu, Y., et al., Richer Convolutional Features for Edge Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: p. 5872-5881.
[10] Zhang, K., et al., Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, 2016. 23: p. 1499-1503.
[11] Cao, Q., et al., VGGFace2: A Dataset for Recognising Faces across Pose and Age. 2018 13th IEEE International Conference on Automatic Face \& Gesture Recognition (FG 2018), 2018: p. 67-74.
[12] Zheng, T., W. Deng and J. Hu, Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments. ArXiv, 2017. abs/1708.08197.
[13] Guo, L. and H. Zhang, A white-box impersonation attack on the FaceID system in the real world. Journal of physics. Conference series, 2020. 1651(1): p. 12037.
[14] Bhatt, R., N. Naik and V. Subramanian, SSIM Compliant Modeling Framework With Denoising and Deblurring Applications. IEEE Transactions on Image Processing, 2021. 30: p. 2611-2626.

AUTHORS

Shuya Tian is currently studying for a master's degree in software engineering from Southwest University of China. Her research interests include adversarial attack, adversarial defense and face recognition.

Xiangwei Lai Ph.D associate professor of computer and information science college Southwest University (SWU), Chongqing, China. The research interests is the security of artificial intelligence, human-computer interaction use affective computing method and affective pattern identification by Physiological signals.