AI security——The application of adversarial samples in face recognition

ZhangLeJun, WuZhiqiang ,ZhangLiFeng, Seiichi SeriKawa

*School of Information Engineering, Yangzhou University, Yangzhou, China

*Electronic Engineering,Kyushu Institute of Technology,kikaKyushu,Japan

Email:,zhanglejun@yzu.edu.cn, edifierwuwu@163.com
zhang@elcs.kyutech.ac.jp, serikawa@elcs.kyutech.ac.jp

Abstract

With the development of economy and the advancement of science and technology, society's requirements for the accuracy of face recognition have gradually increased, and its dependence on face recognition technology has become stronger and stronger. When deploying a neural network in a high-security environment, it is usually easier to ignore the security threat caused by vulnerability. Fully analyze the vulnerability of traditional face recognition technology, and use the generated adversarial samples to design a novel eyeglass patch sample, which can successfully deceive the face recognition system based on convolutional neural network. In addition, the security of artificial intelligence can bring new opportunities for face recognition. Applying artificial intelligence technology to face recognition can effectively improve accuracy. The research purpose of this paper is to analyze the application of adversarial samples in face recognition based on the security of artificial intelligence in order to solve the problems of low security and high vulnerability in traditional face recognition.

Keywords: artificial intelligence; security; adversarial samples; face recognition

1.Introduction

1.1 Literature review

In recent years, a lot of researches on face recognition has been carried out in academia. Zhang Jiasheng and others analyzed the application of adversarial samples in face recognition. They believed that with the increasing reliance

of society on face recognition technology, it is easy to ignore the vulnerability caused by the deployment of neural networks in high-security situations. Security threat. Therefore, designing a novel bright glasses patch sample can successfully deceive a face recognition system based on a convolutional neural network [1]. Ma Long and others studied the application of FLDA in single-sample face recognition, and proposed to superimpose a large number of general samples with every single sample according to a certain ratio, increase the total number of training samples of each class, and can effectively use the FLDA method for features Extraction. Projecting all samples to the feature subspace, and then using the nearest neighbor method to complete face recognition, to a certain extent, alleviate the impact of facial expressions, poses, lighting and other factors on the recognition effect and improve the recognition rate [2]. Xu Tao carried out canonical correlation analysis and its application in face recognition, and considered that canonical correlation analysis (LPCCA) is a new algorithm that can solve a large number of nonlinear problems. It achieves the goal of solving nonlinear problems through a locally linear method. Not only can it maintain the local structural information of each sample set, but it can also obtain the maximum relevant information between two groups of samples. Applying LPCCA to face recognition is very effective [3]. Zhang Ying and Wang Yaonan studied the face recognition algorithm based on the Gabor filter envelope, and proposed a face recognition algorithm based on the two-dimensional Gabor filter envelope. By ignoring the sine part of the Gabor filter and retaining the Gaussian component, a rectangular envelope was constructed. , The elliptic filter is

converted into a rectangular filter, and then feature extraction is performed, so that only the main energy interval of the filter is calculated without affecting the feature extraction performance, and those parts outside the interval are ignored [4]. Liu Xiaojun and others proposed a new face recognition method based on the hidden Markov model. This method uses singular value decomposition to extract facial image features as the observation sequence, which reduces the amount of data storage and calculation, and improves the recognition rate [5]. Summarizing the existing research, it is known that although scholars have richer research on face recognition and have significant research results, but the perspective is relatively single, and few scholars have researched face recognition in combination with the security of artificial intelligence. Therefore, this paper studies the application of artificial intelligence security-adversarial samples in face recognition, which has very important practical significance.

1.2 Purpose of research

After entering the 21st century, face recognition technology has developed very rapidly, and is widely used in all aspects of identity verification, door inspection systems, and personnel monitoring. In recent years, scholars have gradually studied the deep learning of artificial intelligence, and the security of artificial intelligence has caused widespread concern in the society. However, with the advent of adversarial samples, the security of face recognition is threatened. Professor Parham Aarabi of the University of Toronto and his team of graduate students have developed an algorithm that can dynamically disrupt existing face recognition systems. In response to this situation, their solution is to use adversarial training in deep learning technology to allow two artificial intelligence algorithms to confront each other [6]. At present, deep neural networks are widely used in many aspects, such as autonomous vehicles, cancer detection, etc. Researchers need to understand the various ways that this type of model is vulnerable to attack. Generally speaking, in the field of image recognition, adding a small interference to an image can successfully deceive a typical classification network, resulting in image classification errors. This disturbed image is an adversarial sample and can be used to attack the network [7]. Therefore, the research purpose of this paper is

to analyze the application of adversarial samples in face recognition combined with the security of artificial intelligence, which has an important role in promoting social development.

2. Related concepts

2.1 Adversarial sample

The adversarial sample refers to the input sample formed after deliberately adding some small interference to the data set. This situation easily results in that the model will produce incorrect output results although it is under the premise of high confidence. Some scholars have suggested that the reason why adversarial samples occur is because of the existence of excessive linearity [8]. Because neural networks are generally constructed on the basis of linear blocks, in many experiments, the overall function implemented by the neural network will present a highly linear feature. And these linear functions are easier to optimize. However, it should be noted that if a linear function has many inputs, the value will quickly change in some way [9].

2.2 Face recognition

Face recognition is a scientific method for identifying different identities based on facial feature information. Face recognition uses a camera or camera to collect images or videos of human faces, automatically detect faces in images or videos, and then perform corresponding face recognition on detected faces [10].

3. Artificial Intelligence Security Analysis

Artificial intelligence is a comprehensive intelligence that combines human intelligence with natural intelligence. In the 1950s, scientists led by McCarthy, Minsky, Rochester, and Shannon introduced the concept of artificial intelligence. Since then, artificial intelligence has been widely used in robotics, economic and political decision-making, control systems, and simulation systems [11]. With the development of artificial intelligence, artificial intelligence security has become a hot topic that has been constantly discussed in news and technical journals. Generally speaking, artificial intelligence security

refers to the positive or negative effects that technology can have on society.

4. GAN-based adversarial sample generation

Traditional face recognition technology is face recognition based on visible light images. This technology has many problems. If the recognition effect is poor in a low-light environment, it is difficult to meet the needs of current social development. Moreover, in the specific practice process, traditional face recognition technology still has problems such as low technical content and slow response [12]. In view of this, this article mainly proposes GAN-based adversarial sample generation strategies to confuse deep neural networks. The first part analyzes the sample generation, and the second part conducts detailed experimental analysis.

4.1 Sample generation

Generative adversarial network GAN is a deep learning model, mainly composed of a generator (G) and a discriminator (D) [13]. Among them, the generator's role is mainly to learn the real data distribution. With the input variable z , the generator G generates the sample $G(z)$ that obeys the real data distribution as much as possible. The role of the discriminator is to determine whether the input data comes from the generator G or the real data i . If the input comes from the generator G , the label is 0 and the discrimination is false; if the input is not from the generator G , the label is labeled 1, and the discrimination is true.

This article mainly generates a patch of glasses through training to generate an adversarial neural network. This patch can be printed and attached to the frame of the glasses. When the attacker wears it, it is not easy to be detected, but it can disrupt the face recognition system and produce a certain attack effect [14]. In this process, three deep neural networks are mainly trained, namely a generator, a discriminator, and a pre-trained classification function $F(I)$. When inputting x to a deep neural network (DNN), minimizing the generator to generate unobtrusive adversarial output, thereby confusing $F(I)$, the optimization goals are as follows:

$$Loss_G(Z, D) = K \cdot \sum_{z \in Z} Loss_F(i + G(z)) \quad (1)$$

Among them, $Loss_G$, a commonly used loss function for machine learning, is shown in formula (2). By minimizing the loss, the generator can generate real inconspicuous eyeglass image output, and then the discriminator is misled and the discriminator considers the generated glasses An image is a real image. In the process of training the generator, by maximizing the loss, the generated spectacle image is attached to the face image, thereby successfully deceiving the deep neural network.

$$Loss = \max \{ F(i^* | x + r)_y : y \neq y_t \} - F(i^* | i + r)_{y_t} \quad (2)$$

Loss definitions are then made for non-targeted attacks and targeted attacks. It should be noted that the weight of $F(I)$ does not need to be changed during the training process (the main reason is that during the test phase, the expected adversarial output can successfully deceive the same deep neural network).

4.2 Experimental analysis

In order to further analyze the effectiveness of the above adversarial sample generation strategy, the author conducted an experimental analysis on this. First, the data set was collected, and a total of 2,000 related pictures were collected through major websites. Second, set up the experiment. In order to achieve these goals and to be able to perform effective training, the author chose a deep convolution generation adversarial network for analysis. In addition, in order to evaluate the attack performance of the proposed attack strategy, 10 attackers were randomly selected for VGG and OpenFace. As a result, it was found that after using the attack strategy proposed in this article, VGG was successfully attacked and OpenFace was also successfully attacked.

5. Application of adversarial samples in face recognition

At this stage, the application of adversarial samples in face recognition mainly has the following two. First, white-box attacks, that is, obtaining all information and parameters inside the machine learning model to attack, and generating adversarial samples based on the gradient of a

given model, such as Fast Gradient Sign Method (FGSM) [15]. Second, the black box attack, that is, the neural network structure is a black box, and only the model's input and output are used to generate adversarial samples. When the attacker cannot access the model details, the white box attack is not applicable, and the black box attack is needed [16].

Although face recognition technology has achieved great success in various applications, a completely decision-based black box attack—the evolutionary attack method combines dimensionality reduction technology and evolutionary algorithms for pictures, even if the structure and parameters of the model are completely unclear. Under the circumstance, the "best" attack on the identification system can also be implemented without requiring a large number of accesses and being difficult to detect. For example, for face payment in a supermarket, you may only need to wear a pair of special glasses to steal someone's account for payment; the identity authentication system of a station or bank may also misunderstand and steal the identity of others due to simple interference. Of course, the advancement of counterattack technology will also promote the continuous development of defense technology and artificial intelligence security technology. All research aims to create a safe and reliable artificial intelligence.

6. Conclusion

In summary, this paper analyzes the application of artificial intelligence security-adversarial samples in face recognition, and concludes that using a generative adversarial network to design a novel bright glasses patch sample can successfully deceive convolutional neural network-based Face recognition system. With the development of science and technology, in the future, when constructing video recognition models that will be deployed in the public domain in the future, it must be assumed that they will face the threat of false video, especially video confrontation samples, and design corresponding defense measures. On the whole, this paper has certain limitations on the robustness of the adversarial sample. It is suggested that related scholars can apply case and empirical analysis to conduct more in-depth and detailed research in the future.

References

---

1. Zhang Jiasheng, Liu Jianming, Han Lei, et al. Research and application of adversarial sample generation in face recognition [J]. *Journal of Computer Applications and Software*, 2019, 36 (05): 164-170.
2. Ma Long, Wan Kangkang, Han Xiaochun. Application of FLDA in single-sample face recognition [J]. *Journal of Computer Applications and Software*, 2014 (04): 181-183.
3. Xu Tao. Locally maintained canonical correlation analysis and its application in face recognition [D]. Nanjing University of Science and Technology, 2009.
4. Zhang Ying, Wang Yaonan. Face recognition algorithm based on Gabor filter envelope [J]. *Journal of Image and Graphics*, 2008 (12): 69-75.
5. Liu Xiaojun, Wang Dongfeng, Zhang Lifei, et al. A face recognition method based on singular value decomposition and hidden Markov model [J]. *Chinese Journal of Computers*, 2003, 26 (3): 340-344.
6. Chen Fubing, Chen Xiuhong, Zhang Shengliang, et al. Face recognition method based on module 2DPCA [J]. *Journal of Image and Graphics*, 2006, 11 (4): 580-585.
7. Gao Quanyue, Liang Yan, Pan Quan, et al. Problems and solutions of SVD for face recognition [J]. *Journal of Image and Graphics*, 2006, 11 (12): 1784-1791.
8. Yu Bing, Jin Lianfu, Chen Ping. Face Recognition Using Standardized LDA [J]. *Journal of Computer-Aided Design & Computer Graphics*, 2003, 15 (3): 302-306.
9. Zhu Shuxian, Zhang Renjie. Comparison of BP and RBF neural networks in face recognition [J]. *Journal of Instrumentation*, 2007, 28 (2): 375-379.
10. Qing Laiyun, Shan Shiguang, Chen Xilin, et al. Face recognition under arbitrary illumination based on spherical harmonic-based images [J]. *Chinese Journal of Computers*, 2006, 29 (5): 760-768.
11. Wang Shoujue, Qu Yanfeng, Li Weijun, et al. Comparative study of face recognition effect based on bionic pattern recognition and traditional pattern recognition [J]. *Chinese Journal of Electronics*, 2004, 32 (7): 1057-1061.

12. Hong Quan, Chen Songcan, Ni Xuelei. Sub-model canonical correlation analysis and its application in face recognition [J]. Acta Automatica Sinica, 2008, 34 (1): 21-30.
13. Li Xuelong. Artificial Intelligence Theory and Its Application in Face Recognition [J]. Science & Technology Horizon, 2014 (3): 53-53.
14. Yang Zhangjing, Huang Pu, Zhang Fanlong, et al. Centerline neighborhood discrimination embedding algorithm and its application in face recognition [J]. Pattern Recognition and Artificial Intelligence, 2015, 28 (12): 1100-1109.
15. Ma Yukun. Research on Key Technology of Face-based Security Identity Authentication [D]. Beijing University of Technology.
16. Shi Jiarong, Jiao Licheng, Shang Fanhua. Tensor Completion Algorithm and Its Application in Face Recognition [J]. Pattern Recognition and Artificial Intelligence, 2011, 24 (2): 255-261.