Journal on Artificial Intelligence
DOI:10.32604/jai.2021.014175
Article

Tech Science Press

An Adversarial Attack System for Face Recognition

Yuetian Wang, Chuanjing Zhang, Xuxin Liao, Xingang Wang and Zhaoquan Gu*

Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, 510006, China
*Corresponding Author: Zhaoquan Gu. Email: zqgu@gzhu.edu.cn

Received: 04 September 2020; Accepted: 15 March 2021

Abstract: Deep neural networks (DNNs) are widely adopted in daily life and the security problems of DNNs have drawn attention from both scientific researchers and industrial engineers. Many related works show that DNNs are vulnerable to adversarial examples that are generated with subtle perturbation to original images in both digital domain and physical domain. As a most common application of DNNs, face recognition systems are likely to cause serious consequences if they are attacked by the adversarial examples. In this paper, we implement an adversarial attack system for face recognition in both digital domain that generates adversarial face images to fool the recognition system, and physical domain that generates customized glasses to fool the system when a person wears the glasses. Experiments show that our system attacks face recognition systems effectively. Furthermore, our system could misguide the recognition system to identify a person wearing the customized glasses as a certain target. We hope this research could help raise the attention of artificial intelligence security and promote building robust recognition systems.

Keywords: Adversarial attack system; face recognition; physical world

1 Introduction

With the rapid development of computing ability and the advent of the big data era, the artificial neural network has returned to the field of the research on artificial intelligence (AI), and ushered in a high-speed and high-quality development. Many kinds of deep neural networks (DNNs) not only achieved excellent achievements of academic research, but also made great progress in various practical applications, including image processing, natural language processing, and speech recognition. In addition, in many physical information systems, DNN has become an important component in these systems, such as face recognition, intelligent driving, various intelligent voice assistants, spam filtering, etc. The vision system of automatic driving vehicle can do better in identify pedestrians, vehicles and road signs by adopting the DNNs, while face recognition system can be better applied for personnel monitoring and precise access control [1].

However, as the DNNs are adopted in various applications, the DNNs lack intrinsic interpretability, the fragility and security problems of DNNs have been studied in many recent works. Such security problems have been particularly prominent especially when people gradually rely on the usage of DNNs. Many researchers show that DNNs are very vulnerable to adversarial attacks [1–3], which add well-designed antagonistic interference into the original input to generate the adversarial examples. These adversarial examples can mislead the DNNs easily such that the models would work incorrectly [4]. Such adversarial examples generated by the attack methods will bring many serious security problems when DNNs are applied in the real world. In the face recognition systems that are widely used in video surveillance and access control, the adversarial input of confrontation sample could mislead the system to mistakenly identify the violator as a compliance person or himself, which may result in different degrees of consequences, threaten the personal safety of property, and may even bring catastrophic consequences [5].

Considering the related field of adversarial images in the digital domain, there are abundant methods to generate the counter samples and attack DNNs, such as the well-known fast gradient sign method (FGSM) [2], and many works also show that the generated adversarial examples could attack different models with high success rate. However, in the physical domain of the real world, as the attacker can only change himself rather than control the input image of the DNN model, the physical attack against the DNNs will be affected by many environmental factors, such as lighting, posture, angle, etc. This makes the physical attack harder than the image domain attack. Some works transplant the methods of digital domain directly to physical domain, the effectiveness of the attack methods is not necessarily as good as the original attacks [6]. When the attacks are conducted in the physical domain, we also need to consider the concealment of the attack methods. If the added disturbance is too large, we can mislead the face recognition system successfully, but it is too conspicuous and easy to attract people's attention. Then the adversarial attack is meaningless in the physical world. How to conduct effective attacks in the real world has become an important topic. For example, it proposed a method to achieve physical attack in [7] by using a special eyeglass frame to interfere with the face recognition system.

Relying on the previous works, it is obvious that the attack methods can be utilized to attack real face recognition systems in the physical domain, which implies the attackers do not need to modify the input image to the system. By designing customized glasses, the person wearing the glasses could mislead the face recognition system such that he or she is identified as a wrong person. Based on the system, tests and analysis can be conducted to evaluate the defensive performance of different face recognition systems. By mining vulnerability and examining the check points of the face recognition systems, the robustness of the recognition systems against the physical domain attacks could be promoted, and thus enhance the security level of the related applications.

In this paper, we first briefly introduce the adversarial samples, a security problem that exists against DNNs. Then we explain some existing attack methods that could generate adversarial example efficiently. After that, we describe the physical attack system we designed against face recognition from the aspects of system design, framework, algorithms, and the system functions. The effects of the related algorithms are shown in this system. Finally, we shortly discuss and summarize some possible functions that could be added in the future.

2 Related Work

2.1 Adversarial Examples

Adversarial examples are the generated samples in which database are intentionally perturbation-employed. The generated adversarial examples will lead to a wrong export of the model at high confidence. Many works show that the intentionally fabricated samples could achieve very high attack success rate, by perturbing the optimization process of the models. In addition, the generated samples are quite familiar with the original inputs, which causes humans cannot be aware of the difference between the original examples and the adversarial examples easily, while the DNNs may output a totally diverse prediction [4]. Some works also try to explore the essence of the adversarial examples, and it is generally assumed that there are two reasons that may lead to the adversarial examples. On the one hand, since the input examples for training the model cannot cover all possible situations, there would exist some extreme cases that might evade the model. On the other hand, since the DNNs lace theoretical explanation and many DNNs show linear character in high dimensional space, these models have some intrinsic shortages that might be attacked with such intentional crafted samples [3].

2.2 Adversarial Examples in Digital Domain

Many methods that can generate adversarial examples have been proposed, and most of them are adopted to attack white-box models, where the attackers have full access to the model [8], such as the training data, the architecture, the parameters, etc. One pioneering work, fast gradient sign method (FGSM), is proposed in [2], which fabricated adversarial examples using the first-order approximation of the loss function. After that, the method to anti-infer target attacks is introduced in [9], which works on the basis of optimization method. To be concrete, these attacks create a target function, and the goal is to maximize the difference between the ground-truth label and an incorrect label that the attacker expects. Meanwhile, the function also tries to minimize the similarity between the original input and the generated sample. Recent works have shown that digital adversarial examples show good transferability, as the generated adversarial examples against a white box model could also attack a black box mode (with no access to the model) with high probability. The transferability implies that it is possible to generate adversarial examples to attack some real DNNs [10].

2.3 Adversarial Examples in Physical Domain

There are many effective methods in generating adversarial examples in the digital domain, but these methods might not work well in the physical domain. In the digital domain, the attackers could modify the input image, but they cannot access the image in the physical domain. On the contrary, the attackers can only modify the physical environment or the objects in the environment, the cameras would capture the current image and it is sent to the DNNs directly. Controversy about the effectiveness and feasibility of the methods has been drew when these methods are transferred directly in the physical domain. It is supposed that the existing anti-infer method used for halting detection could only be available under carefully chosen circumstances in [11]. In addition, in most real cases, there is no need to concern about the adversarial examples because a well-trained network could detect the adversarial example from several distances and angles easily. Nevertheless, the contrary opinions exist and they show the adversarial examples in the nature physical domain. The adversarial examples are printed out in [12], and then the smart-phone camera could still classify the captured images incorrectly. Some customized images are printed on the spectacles to attack face recognition system in [7]. These works have proved that the change of the gesture, light, angle and the distance between the camera would cause tiny effect during the attack and the adversarial examples might work in a relatively stable physical condition.

As the face recognition systems have been widely adopted in current lives, especially in many sensitive areas which deeply related to individual properties and social safety. It has been more and more important to enhance the safety of the face recognition systems such that they can resist various attacks. In this paper, we implemented an attack method against face recognition system by wearing specific eyeglasses, and we hope this would improve people's attention in enhancing these systems.

3 System Design

Before we implement the physical domain attack system, we deployed a web application to verify out attack method for the physical model and this is a prerequisite for finally applying the system to the practical application in the real world.

Consider the actual scenario in the physical model that the attacker cannot tamper with the input, it is impossible to access the digital information directly which is captured from the physical world. We assume the attacker can only launch the attack by changing features of the person, such as changing the accessories the person wears. In addition, the attacks in the physical world should not be obvious to attract people's attention. Hence, in our work, we design a specific glass frame to conduct the attack and we show the attack performance in our deployed application.

3.1 Framework

The physical attack system adopts the B/S mode. The distributed feature of the B/S mode has many advantages, such as it allows easy extension of business logic, it is high-efficiency for early-stage development, it can be concurrent developed by distributed git version control, and it has low maintenance cost for post-stage development. To alter or add certain business logic, only a small part of the web application needs to be modified accordingly.

The background server is implemented by Flask, a lightweight architecture developed in Python suiting for agile software development. As our physical attack algorithm is also implemented in Python, such architecture facilitates better interaction between the client and server. The agile development combination of Bootstrap and jquery are also adopted in the web application.

3.2 Prototype Design

We need to design a prototype of the system as Fig. 1 before we develop the application. This prototype needs to meet our existing requirements and can be updated iteratively

Before starting to build the attack system, we need to clarify the functions of this system, design a functional layout, and make reasonable arrangements while noticing the scalability of subsequent new functions.

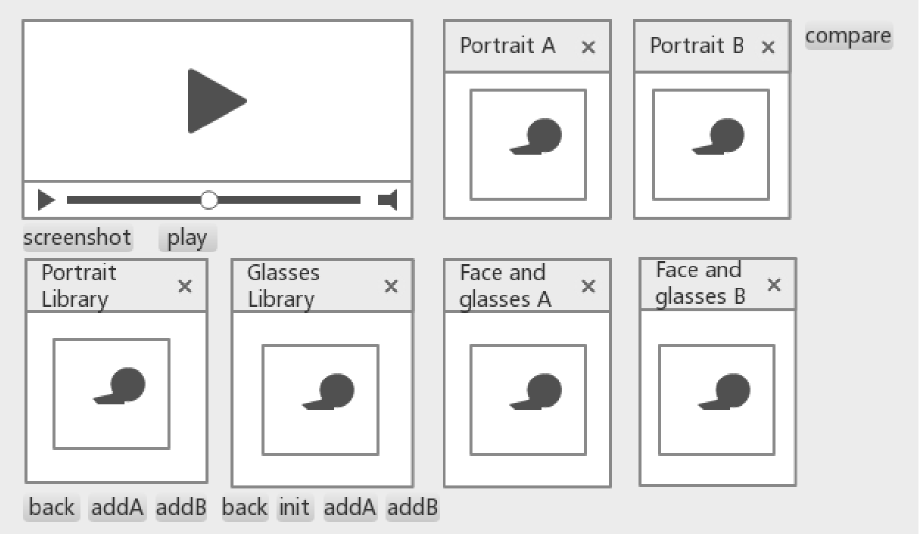


Figure 1: Prototype design of this system in this article

3.3 System Functions

3.3.1 Real-Time Video Reading

The real-time video reading function is design as the top left corner in the figure, which can capture the video frames from the camera. In addition, it can capture the snapshot through real-time video streaming. The function displays the real-time videos timely.

3.3.2 Human Face Recognition

The function judges whether the captured person is the same person after it gains the image through the captured snapshots.

In our work, the deep convolutional neural network VGG16 trained VGG Face model is used as the face recognition algorithm. The input size of the captured image is 224*224, and the output of the model is a vector with 512 dimensions. The vector contains a lot of face features which can be used as the main criterion to decide whether two different images have the same person. This recognition can be achieved by comparing the cosine distance of the output vector of the two images.

In the preprocess operations, 68 key check points are identified by importing the model provided by dlib in Python. After face alignment, the 68 check points of all images are basically distributed around identical locations which facilitate the attacking procedure.

After the face alignment, we can find the suitable position for adding the glass frame and the shape of the glasses can also be designed easily.

3.3.3 Generate Attack Glasses

The function generates different glasses by the attack method. In order to implement the function, the generated glass frame shape is close to a normal one. The color of the generated glass frame is randomly generated and smoothed, considering column traverse, row traverse and randomness. As verified in the experiment, the randomly generated colored glasses, despite lacking any insight of the recognition model, can interfere the output of VGG Face model, by outputting incorrect result.

3.3.4 Simulate Physical Attack

After randomly choosing one of the generated glasses patterns, the system would automatically add this glass frame to the face collected by the camera. In that case, the face recognition algorithm in the system cannot recognize the generated image correctly, which realizes the attack on the face recognition system.

3.3.5 Generate Real Attack Glasses

The glasses patterns generated by the attack algorithm can also be printed out and the real attack glasses can be designed. Then the users can attack the face recognition system after wearing the glasses and the system would not identify the users correctly with high probability.

4 System Performance

4.1 Algorithm Performance

We discuss the results of experiment here, as shown in Tab. 1. We use the VGG Face model for recognition and three persons (we denote them as Subject A, B and C, respectively) participated in the experiment. The average difference of the model's output between the original images of a same person are less than 0.5. However, the difference value increased significantly when we add different special glasses to the original images. As shown in Tab. 1, we present the average difference values between the images with three kinds of glasses (Glasses 1, 2, and 3, respectively) and the original images. The recognition results of the generated images are also presented in the table. It is obvious that three persons are recognized incorrectly when the glasses are added.

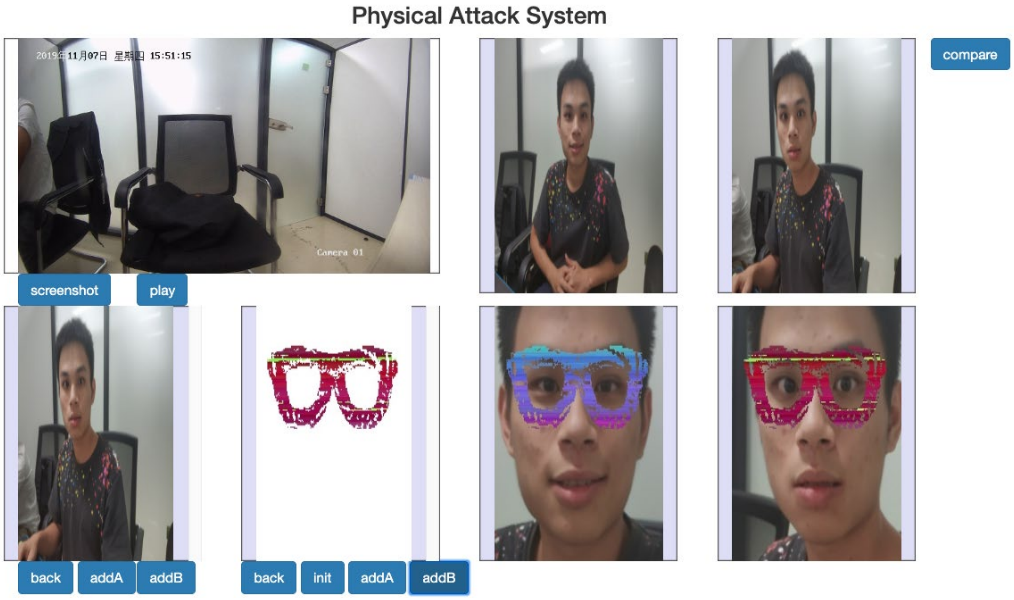
Table 1: Average difference and recognition result after wearing glasses

	Subject A		Subject B		Subject C	
	Difference value	Recognition result	Difference value	Recognition result	Difference value	Recognition result
Glasses 1	2.56	Not A	3.11	Not B	1.93	Not C
Glasses 2	3.18	Not A	3.37	Not B	2.44	Not C
Glasses 3	3.46	Not A	3.80	Not B	2.86	Not C

4.2 System Performance

The interface of the system mainly contains several functional modules such as real-time video window, portrait capture, glasses selection, adding glasses, etc. We present the whole interface of the system as Fig. 2 and we introduce its functional modules in details. There are some different parts on the system interface and they represent different functional modules.

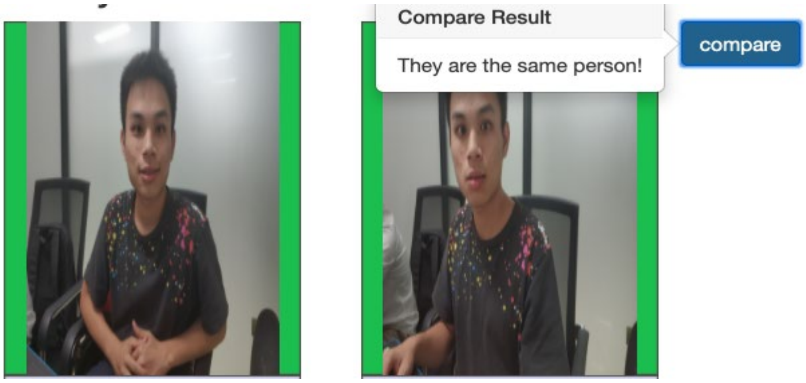
In Fig. 2, the top left part shows a real-time video stream obtained from the camera, and we can capture the face images in this functional module. In the lower left part, the user can select different specific face images and different specific glasses, respectively. The selected images and glasses would be displayed and applied in the attack. The top right part shows the two images the user selected before, while the lower right part shows the face images that are added with the selected glasses. User can choose any two of the four images on the right part to compare whether they are the same person, and the system will output the recognition result.



Physical Attack System

Figure 2: System interface

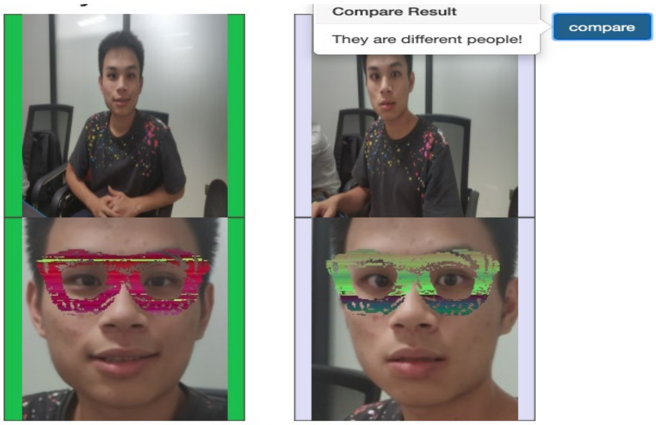
For example, if the user selects the two images in the top right parts, they are the same person and the system would output the recognition result as Fig. 3. However, if the user selects one image in the top right part and one image in the lower left part as Fig. 4, the system would identify the person wearing the glasses as another person, which confirmed the adversarial attacks with the glasses.



Compare Result

They are the same person!

Figure 3: The system recognizes that they are the same person



Compare Result

They are different people!

Figure 4: The system recognizes that they are different persons after wearing glasses

5 Discussion

The implemented system in this paper can accomplish basic image capture and face comparison functions, and it can attack the recognition system by adding different glass frames to the victim person. By printing the glass frames in the real world, the tester wearing the glass frame cannot be identified correctly by the face recognition system. During the process of the optimization and generalization process, there are still some shortages in mapping the glass frames in the image domain to the real glasses in the physical domain. Considering the RGB mode of the picture's generation and display, it may not able to be printed the exact colors by the printer, which implies color differences exist between the generated glasses images and the printed glasses. An appropriate approach is to print out all the color the printer can offer and compare the printable colors with the captured colors by the camera. Since there are still defects considering the environmental light and camera angle when capturing the images, there might still need some other procedures to handle the problem, such as homogenization and smoothing.

6 Conclusion

In this paper, we implemented an attack system against the VGG Face model, which can identify the images correctly. In our system, we can clearly get the successful effect of the attack on VGG Face with special glass, and also control the production of different special glasses. We introduce the procedure of generating the glass frames against the white-box model, and this might be applied to attack some real black-box model as well. We introduced the system design and the main functions we accomplished, and this system could be applied to verify the attack performance against the face recognition model. This effect can be achieved by using different glasses to attack the detected face recognition system and comparing the success rate.

Based on existing work, we need to consider the impact of the printer generated color and ambient light on the production of glasses in the future. This production may reduce the difficulty of attacking the face recognition system. Apart from drawing the attention about the security problems of the face recognition systems, it would be an important and interesting work in the future to protect people's privacy, while reserving the key features for efficient face recognition. Our system may promote the improvement of face recognition system in security.

Acknowledgement: We thank the editor and the anonymous reviewers for their helpful comments in improving the paper.

Funding Statement: This work is supported in part by the National Natural Science Foundation of China under Grant 61902082, U1636215, and the Guangdong Province Key research and Development Plan under Grant 2019B010136003.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan *et al.,* "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[2] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd Int. Conf. on Learning Representations*, Lille, France, 2015.

[3] J. Kos, I. Fischer and D. Song, "Adversarial examples for generative models," in *2018 IEEE Security and Privacy Workshops*, San Francisco, CA, USA, pp. 36–42, 2018.

[4] T. Miyat, S. Maeda, M. Koyama, K. Nakae and S. Ishii, "Distributional smoothing with virtual adversarial training," arXiv preprint arXiv:1507.00677, 2015.

[5] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *11th ACM/IEEE Int. Conf. on Human-Robot Interaction,* Christchurch, New Zealand, pp. 101–108, 2016.

[6] J. Lu, H. Sibai, E. Fabry and D. Forsyth, "No need to worry about adversarial examples in object detection in autonomous vehicles," arXiv preprint arXiv:1707.03501, 2017.

[7] M. Sharif, S. Bhagavatula, L. Bauer and M. K. Reiter, "Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition," in *ACM Conf. on Computer and Communications Security*, New York, NY, USA, pp. 1528–1540, 2016.

[8] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: a survey," *IEEE Access,* vol. 6, pp. 14410–14430, 2018.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural network," in *38th IEEE Sym. on Security and Privacy,* San Jose, CA, USA, pp. 39–57, 2017.

[10] N. Papernot, P. Mcdaniel and I. J. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," arXiv preprint arXiv:1605.07277, 2016.

[11] A. Kurakin, I. J. Goodfellow and S. Bengio, "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, 2016.

[12] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati *et al.,* "Robust physical-world attacks on deep learning visual classification," in *IEEE Conf. on Computer Vision and Pattern Recognition,* Salt Lake City, Utah, USA, pp. 1625–1634, 2018.