JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Attack Analysis of Face Recognition Authentication Systems Using Basic Iterative Method (BIM)

[1]Supantha Sarker: supanthasarker@gmail.com, [2]Md Muntasir Mamun Siam: siammuntasir588@gmail.com, [3]Md Shayok Ul Islam: shayok11800031@gmail.com,[4]Md Asibur Rahman Akash: text.2.mine@gmail.com, [5]Md Monir Hossain: hossainkhan1164095@gmail.com, Gagandeep Kaur: gagandeep.23625@lpu.co.in

School of Computer Science Engineering

Lovely Professional University, Punjab, India

1. ABSTRACT

Machine Learning models are used for a variety of tasks, such as picture categorization, virus detection, and network intrusion detection. However, subsequent studies have shown that even the most advanced deep neural networks, which excel at such tasks, are subject to a bug. Adversarial Examples are a type of harmful input. These are non-random inputs that are nearly indistinguishable from natural data but are incorrectly categorized. In this paper, we attempt to investigate the presence of adversarial instances and highlight some of the numerous approaches developed to exploit deep neural network flaws. over time and give an analysis of such attacks as a subset of ImageNet's visually distinguishable classifications.

Keyword: - Adversarial Examples, virus detection, Machine Learning, deep neural networks.

2. INTRODUCTION

The Basic Iterative Method(BIM) is a FGSM extension. The FGSM is repeated numerous times with a short step size for this attacks. The result is cropped after each iteration to guarantee the perturbation is inside the neighborhood of the original picture.[1] These enhancements enabled the BIM to be a more powerful attack with fewer disturbances. The projected gradient attack, introduced by Madry is the strongest attack based on local first order knowledge about the network (PGD). The attack is similar to BIM, except that the PGD is resumed from a variety of random positions across the 1-inf ball to better comprehend the loss landscape.

Deep learning applications have been used in every part of life as result of a global study in academia and industry. From smart bome gadgets like Amazon echo, Google Home, and Facebook portal to industrial use like drone delivery, warehouse automation, medical imaging and more, there is something for everyone vehicles that drive themselves.[2] These gadgets have been around since the begimmimg of time, in both personal and industrial settings. Deep learning advance have hastened this process. Perception to intelligence transformation quicker and more accurate picture identification allows for real-time reactions and actions models.[3] Smartphones, for example, employ facial detection and recognition to verify that the user is who they say they are.

However, as IoT devices become more widely used, the systems become more vulnerable to a variety of threats. Adversarial examples are one such weakness. These hostile examples are deliberately produced inputs that are intended to deceive the model and reduce its accuracy and real-world performance. These assaults are difficult to detect since they are normally undetectable to humans, but they can significantly reduce the accuracy of the model. The adversaries are asymmetric and are designed to undermine the integrity of deep learning models in certain ways. These issues have plagued deep learning implementation in safety-critical applications such as home security, medical imaging, and autonomous vehicles.

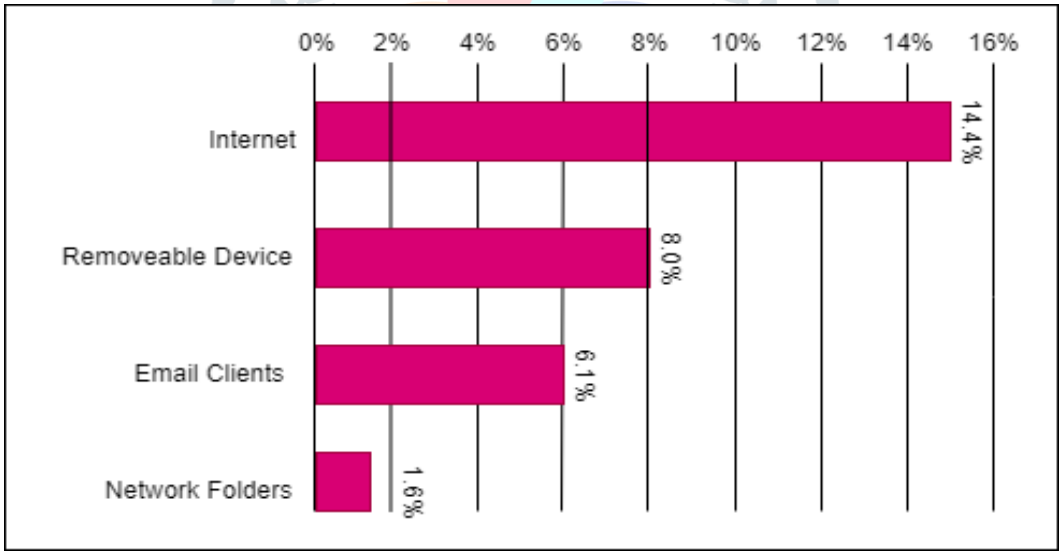

Figure 1. Threats for biometric data processing and storage systems (Kaspersky 2019).

To combat adversarial vulnerability, a significant amount of research was made into developing a variety of defensive measures for adversarial assaults, which may be loosely divided into two methodologies: rectification and adversarial input detection.[4] The purpose of rectification approaches is to recover the model's intended output by strengthening the system's resilience, such as by attempting to eliminate adversarial disruption from the input or by raising the robustness of the model itself. Adversarial detection, on the other hand, seeks to identify an assault that has already happened by evaluating the model's behavior

and communicating aberrant occurrences. While most adversarial detection techniques are evaluated on tiny or low-resolution benchmarks, the goal of this research is to evaluate one of the aforementioned training-manifold-based adversarial detection approaches in a realistic.

Deep learning classifier are rarely employed in facial recognition system; instead, a similarity-based approach is utilized: deep models are used to exact features from visual facial data, and choices are measures among those features.[5] One of the most significant issues in computer vision is face recognition. Since the early 1990s, when the Eigen faces technique was presented, this topic has piqued the scientific community's interest since 2012, DL models, perticularly those using the features of Deep Convolutional Neural Network, have domiated this field, achieving a performance of up to 99.80% and so surpassing human performance on this challenge.[6] Despite the work put into developing extremely strong DL models, such system neverthless have flaws. It has been proven, for example, that when evaluated against low-quality photos, state-of-the-art face classifiers perform poorly.

3. Literature review

Because the face is regarded as the most significant portion of the human body, it plays a vital function in identifying and authenticating a person, and as a result, it may be utilized for a variety of purposes in everyday life. Face recognition systems are becoming increasingly popular across the world, according to (Zulfiqar et al. 2019). They provide secure and dependable security solutions. It is the quickest biometric technology for identifying a person without having to interact with him. It automatically captures the picture of a person from a particular distance or grabs a sample from a video, analyses that image, and recognizes that person, so there is no unnecessary weight. It can be shown by (Mori, Matsugu, and Suzuki 2005) and (Li et al. 2020) that ML approaches, notably artificial neural networks, are employed to create a system that fulfills the function of facial recognition. Convolutional neural networks (CNN) are neural networks that were built expressly for such tasks and have demonstrated efficiency in doing them when it comes to image processing, analysis, and comparison (Albawi, Mohammed, and Al-Zawi 2017). In the case of our research, this form of the network was also utilized.

4. ADVERSARIAL ATTACK:

The key characteristics of adversarial example generating methods are discussed in this section.

4.1 Adversarial Capacity: The amount of knowledge the attackers might obtain about the model determines adversarial capability. According to the attack's capabilities, threat models in deep FR systems are divided into the following groups.

4.2 White-box attack: Expects a thorough understanding of the target model, including its parameters, architecture, training technique, and, in certain circumstances, training data.[7]

4.3 Black-box attack: Without knowing the target model, feeds hostile instances developed during testing to a target model (e.g., its training procedure or its architecture, re or its parameters). Even if the attackers do not have access to the model's knowledge, they can interact with it by using adversarial examples that can be sent.[7]

4.4 Adversarial Specificity: The capacity of an assault to allow a precise intrusion/disruption or because broad mayhem is known as adversarial specificity. According to the specificity of the assault, threat models in deep FR systems may be classified into the following groups.[8]

4.5 Targeted attack: Deceives a model for predicting the wrong label for an adversarial case. This is accomplished in an FR or biometric system by impersonating notable persons.

4.6 Untargeted attack: As long as the results aren't the proper labels, predict the labels of the hostile samples. Face dodging is used in an FR/biometric system to do this.[9] Because it has more options and room to change the output, a non-targeted assault is easier to implement than a focused attack.

5. BASIC ITERATIVE METHOD (BIM)

The Fast Gradient Sign Method is easily extended by the Basic Iterative Method. Instead of one large step, like in FGSM, we take several smaller ones.

$$Y\ adv_0 = X,\ X^{adv}{}_{N+1} = Clip_{x,\varepsilon}\{ X^{adv}{}_N + \alpha sign(\nabla_x J(X^{adv}{}_N, y_{true}))\}$$

At each cycle, the picture pixel values are also trimmed to guarantee that they are in the - neighborhood of the original image.[10] It is assumed that the number of iterations will be min (α + 4, 1.25). This was decided based on a heuristic; it is adequate for the adversarial example to reach the max-norm ball's edge while remaining constrained enough to keep the experiment's computing cost affordable.[14]

BIM was performed on an image of the hen. On adding some perturbation iteratively, we see the misclassification by our model in the top-5 predictions. Here we used = 0.7 and α = 0.2

5.1 Untargeted BIM

According to (Kurakin 2017), iteratively performing FGSM on a sample with a reduced step size leads in a stronger adversarial sample.[5] The output is clipped after each iteration to guarantee that the adversarial sample is within the ε - neighborhood of the original input Xi.

Untargeted adversarial sample Xiadv is claculated using equations

$$X^i_{adv,0} = X^i$$

$$X^i_{adv,N}=\min\{X^i + \varepsilon^i{}_{max}, \max\{X^i_{adv,N-1}+ \alpha \text{sign}(\nabla_x L(X^i_{adv,N-1}, Y^i)), X^i - \varepsilon^i{}_{max}\}\}$$

5.1 Targeted BIM

Both of the strategies we've covered are untargeted, in the sense

that they raise the cost of the proper class without specifying the misclassification target. Now we'll look at a method that looks for misclassifications in a certain class. This would be the target of a targeted attack. This strategy is useful for datasets with a large number of classes because datasets like MNIST have a smaller number of classes but a higher degree of class differentiation. We can acquire uninteresting misclassifications like a change in the animal breed as a misclassification if we perform untargeted assaults on ImageNet. On the other side, it would be fascinating to witness a spider misclassified as a towel.[11]

This technique attempts to misclassify the original picture into the class with the least probability based on the original image's forecast. This class is found by:

$$yLL = \text{argmin}\{p(y|X)\}y$$

To achieve this misclassification, we try to maximize logp(yLL|X) by making iterative steps in the direction of sign {∇X log p(yLL|X)}. This expression equals sign {−∇X J(X, yLL)} for networks with cross entropy loss. Therefore, we have the following method:

$$X^{adv}{}_0 = X,\ X^{adv}{}_{N+1} = \text{Clip}_{x,\varepsilon}\{X^{adv}{}_N - \alpha\ \text{sign}(\nabla_x J(X^{adv}{}_N, yLL))\}$$

The alpha value and number of iterations are the same as in the Basic Iterative Method.[12]

Because the above methods are not exact, we cannot guarantee that we will always have the equivalent adversarial example.[15] We compare the misclassifications per approach vs. ε values.

6. Experimental Setup

Because the attack necessitates the creation and training of a model using machine learning techniques via artificial neural networks, and the convolutional neural network (CNN) is the best network for training models aimed at classifying images, recognizing images, or even detecting objects, the steps to achieve training should be devised.

7. Create a Model for Authentication

7.1 Data Connection

It is required to have a huge number of data, in our instance photos of different people, to train the artificial neural network, the model, and to have the proper relevant content from which the model may learn. The dataset's size is determined by the classification job and the model that has been created. CNN models require large datasets, however, because we apply transfer learning, the dataset may be considerably smaller. We utilized the public "5 Celebrity Faces Dataset"1 DanB (2017), which comprises photographs of five distinct people.[6] This is a simple database for experimenting with computer vision methods. In the files, there are 14–20 images for each individual, organized in folders, but in the assessment files, there are 14–20 pictures for each person.

two main techniques that are used during data collection and preparation are:

- Data processing is the process of applying modifications to data before it can be utilized the act of converting raw data into a clean data collection is known as data processing. To put it another way, anytime data is collected, gathered from several sources It is gathered in unprocessed forms, which is not ideal. The analysis is possible.

- The technique of adding data by changing it is known as data augmentation. existing data without having to obtain fresh data This procedure is quite time-consuming. In DL, this is helpful since there are times when a huge amount of data must be collected. It is quite efficient to add a large quantity of data in this manner. It makes it possible. improved learning as a result of the increased training dataset and the ability to learn from different states, you'll need to use an algorithm. We employed the following main operations in this process: scaling (ensures that the input is in the range [0, 1]), rotation, zoom, shift, and flip.

7.2 Defining Model Architecture

To achieve the most efficient training of the model, the architecture of the model should be specified by identifying the layers to which the data should be transferred. The number and kind of layers, the number of neurons, and the type of activation functions are all unique to any neural network model. These attributes must be assessed and allocated for the model's performance to be satisfactory and acceptable for the role it is to execute. The type of neural network that has been elaborated and implemented is CNN employing the convolution operation, due to the topic's restriction in assessing assaults on the authentication process based on face recognition features.[13] The benefit of CNN is self-evident, given that it is dedicated to challenges involving computer vision and our problems.

VGGNet is the CNN network design suggested by Karen Simonyan and Andrew Zisserman from Oxford University in 2014, as reported in Simonyan and Zisserman (2014). RGB pictures of 224 224 dimensions are used as inputs to the VGG network. The VGG16 architecture contains three completely linked layers and thirteen component layers, whereas the VGG19 architecture has sixteen. Instead of huge filters, there are little filters with three dimensions but greater depth. Two fully linked layers contain 4096 channels in both types of architecture, whereas the third layer has 1000 channels to predict 1000 labels. The categorization is carried out by the last fully connected layer, which employs the softmax activation function. In ImageNet, we utilized the VGG16 model with pre-trained weights as the foundation model for our trained model. Both picture formats with color channel dimensions of "channels - first" and "channels last" are supported by this model.

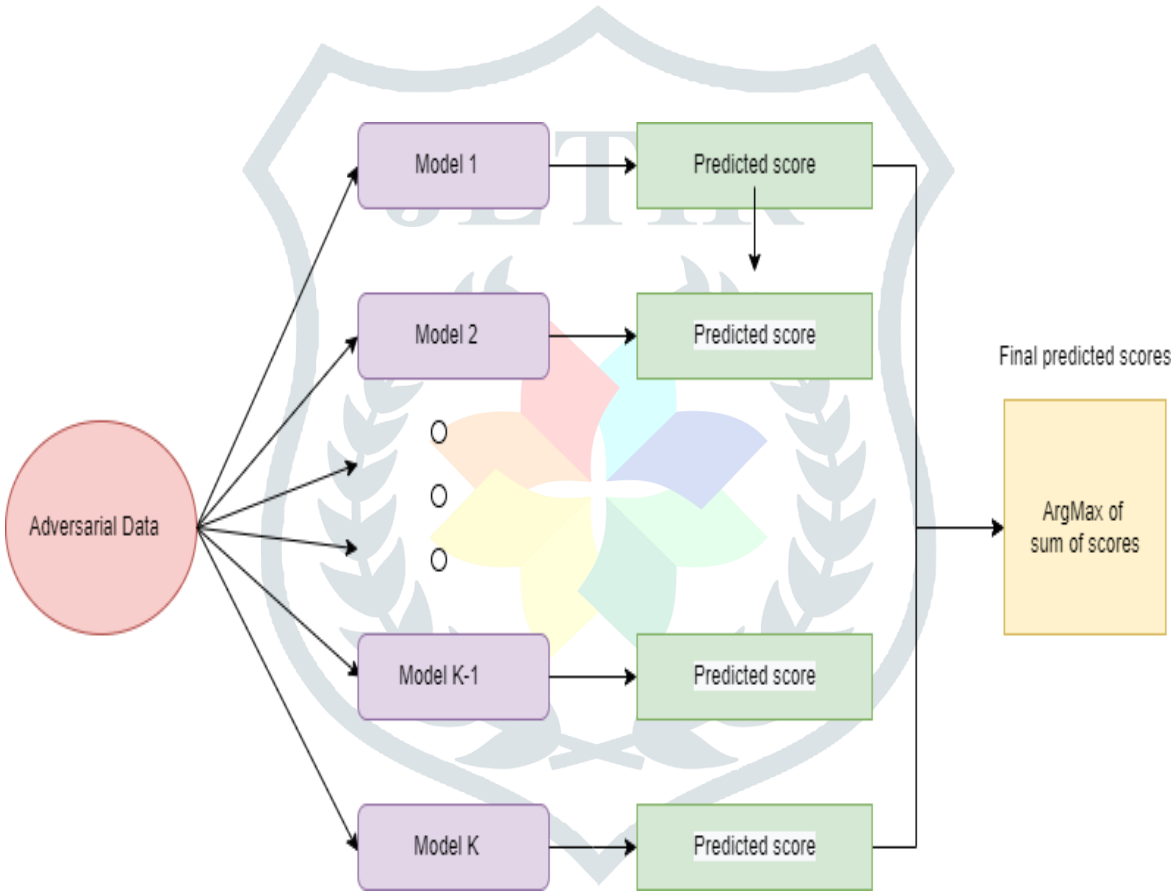

Figure 2: Iteration Model Architecture

7.3 BIM Implementation

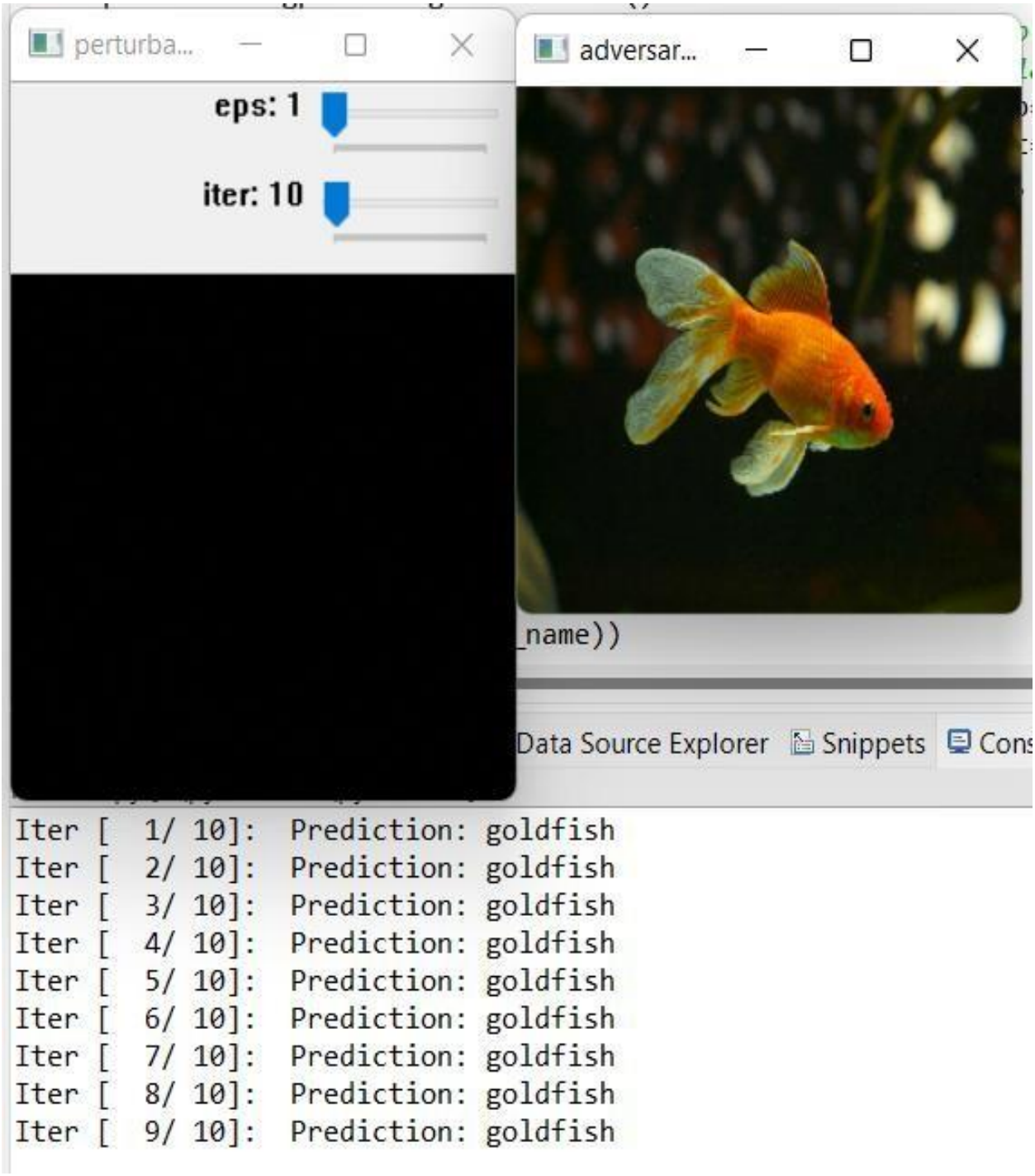

Figure 3: BIM Implementation

8. Result

This project has executed the Basic Iterative Method successfully. The attack is done according to the eps and iteration value given by the user. The project successfully defines the prediction of the image through the class give in the source code.

The user can conduct multiple round of attacks with in the same execution just by pressing the space button. The user has to give eps and iteration value of their choice or else the program will execute another round of attack with the default value. To stop the execution of the program the user just has to press escape. It will close the execution window automatically.

9. Conclusion

The research paper proposes the Basic Iterative Method is to conduct adversial attacks. The program successfully executes and predicts the image correctly. It predicts the image from the class program given in the source code. It can predict any image updated to it. Other types of adversial attacks can also be concluded such Fgsm, Gat, one pixel attacks with some modification in the program. Through the experiments done with the code it seems that Basic Iterative Method is a good way of conducting adversial attacks as the image is successfully predicted. But there is always scope for improvement in this project.

10. References

[1] Zulfiqar, M., F. Syed, M. Khan, and K. Khurshid. 2019. Deep face recognition for biometric authentication. In 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), pp. 1-6, doi: 10.1109/ICECCE47252.2019.8940725.

[2] Mori, K., M. Matsugu, and T. Suzuki. 2005. Face recognition using SVM fed with intermediate output of CNN for face detection. In Proceedings of the IAPR Conference on Machine Vision Applications (IAPR MVA 2005), 410–13. Tsukuba Science City, Japan, May 16-18

[3] Li, Y., Z. Wang, Y. Li, X. Zhao, and H. Huang. 2020. Design of face recognition system based on cnn. Journal of Physics: Conference Series 1601:052011. August

[4] Albawi, S., T. A. Mohammed, and S. Al-Zawi. 2017. Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET), 1–6.

[5] Alexey Kurakin, Ian J. Goodfellow, Samy Bengio, “Adversarial examples in the physical world,” International Conference on Learning Representations, 2017.

[6] DanB. 2017. 5 celebrity faces dataset. https://www.kaggle.com/dansbecker/5-celebrity-facesdataset

[7] Simonyan, K., and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556. September.

[8] Wenqing Liu, Miaojing Shi, Teddy Furon, Li Li. Defending Adversarial Examples via DNN Bottleneck Reinforcement. ACM Multimedia Conference 2020, Oct 2020, Seattle, United States. pp.1930-1938, ff10.1145/3394171.3413825ff. ffhal-02912189f

[9] Massoli, F. V., Carrara, F., Amato, G., & Falchi, F. (2021). Detection of Face Recognition Adversarial Attacks. Computer Vision and Image Understanding, 202(January). https://doi.org/10.1016/j.cviu.2020.103103

[10] Liu, W., Shi, M., Furon, T., Li, L., Liu, W., Shi, M., … Furon, T. (2020). Defending Adversarial Examples via DNN Bottleneck Reinforcement To cite this version : HAL Id : hal-02912189 Defending Adversarial Examples via DNN Bottleneck Reinforcement.

[11] Vakhshiteh, F., Nickabadi, A., & Ramachandra, R. (2021). Adversarial Attacks against Face Recognition: A Comprehensive Study. IEEE Access, 9, 92735–92756. https://doi.org/10.1109/ACCESS.2021.3092646

[12] Nguyen, D. L., Arora, S. S., Wu, Y., & Yang, H. (2020). Adversarial light projection attacks on face recognition systems: A feasibility study. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2020-June, 3548–3556. https://doi.org/10.1109/CVPRW50498.2020.00415

[13] Theagarajan, R., & Bhanu, B. (2020). Defending black box facial recognition classifiers against adversarial attacks. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2020-June, 3537–3547.

https://doi.org/10.1109/CVPRW50498.2020.00414

[14] Kwon, H., Kwon, O., Yoon, H., & Park, K. W. (2019). Face Friend-Safe Adversarial Example on Face Recognition System. International Conference on Ubiquitous and Future Networks, ICUFN, 2019-July, 547–551. https://doi.org/10.1109/ICUFN.2019.8806124

[15]Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., & Busch, C. (2019). Face Recognition Systems under Morphing Attacks: A Survey. IEEE Access, 7, 23012–23026. https://doi.org/10.1109/ACCESS.2019.2899367