International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 02 | Feb 2023 www.irjet.net p-ISSN: 23950072

Adversarial Impersonation: A Study of Mask-Based Attacks on Face Recognition Models

Vaishnavi Bisen[1], Rutuja Farkade[2], Vaishnavi Datir[3], Pragati Lanjewar[4], Janvi Shah[5], Sanyukta Sadhankar[6]

[1,2,3,4,5] Final year, Department of Computer Science & Engineering, Sipna College of Engineering and Technology, Amravati, Maharashtra, India.
[6] Final year, Department of Information Technology, Sipna College of Engineering and Technology, Amravati, Maharashtra, India.

Abstract - *Face recognition (FR) systems have demonstrated reliable verifification performance, suggesting suitability for real-world applications ranging from photo tagging in social media to automated border control. In an advanced FR system with deep learning-based architecture, however, promoting the recognition effificiency alone is not suffificient, and the system should also withstand potential kinds of attacks. In this paper we are trying to figure out accuracy of face recognition model in the absence of adversarial effect and in the presence of adversarial effect. For this two different algorithms that are FisherFace Recognizer and LBPHFACE Recognizer can be used. But in this paper we have used LBPHFACE Recognizer to calculate the accuracy of model with and without adversarial attack.*

Key Words: Adversarial Attack, Face Recognition, Face Mask, FisherFace, LBPH.

1. INTRODUCTION

In this paper we are trying to figure out accuracy of face recognition model in the absence of adversarial effect and in the presence of adversarial effect. For this two different algorithms that are FisherFace Recognizer and LBPHFACE Recognizer can be used. But in this paper we have used LBPHFACE Recognizer to calculate the accuracy of model with and without adversarial attack. We are the first to present a physical universal adversarial attack that fools FR models, i.e., we craft a single perturbation that causes the FR model to falsely classify all potential attackers as unknown identities, even under diverse conditions (angles, scales, etc.) in a real-world environment In the digital domain, we study the transferability of our attack across different model architectures and datasets. We present a fully differentiable novel digital masking method that can accurately place any kind of mask on any face, regardless of the position of the head. This method can be used for other computer-vision tasks (e.g., training masked-face detection models). We craft an inconspicuous pattern that "continues" the contour of the face, allowing a potential attacker to easily blend in with a crowd without raising an alarm, given the variety and widespread use of face masks during the COVID 19 pandemic. We propose various countermeasures that can be used during the FR model training and inference phases.

2. Related Work

2.1 Adversarial Attacks

A technique for finding a perturbation that alters a machine learning model's prediction is known as an adversarial attack. The disturbance may be quite insignificant to human perception[1].The amount of information the attackers have about the model determines their adversarial capacity. Based on the attack's capability, threat models in deep FR systems are divided into the following categories:

The white-box attack assumes complete comprehension of the target model's parameters, architecture, training method, and, in some instances, training data.

Black-box attack uses adversarial examples created during testing without knowing the model's parameters, architecture, or training procedure to feed the target model. Although the attackers do not have access to the model's knowledge, they are able to interact with it through the transferability of adversarial examples[2].

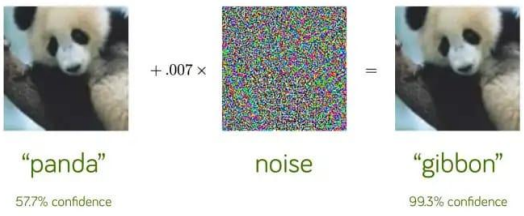

Fig-1: Visualization of original face image (first column), adversarial noise vector of VGG-16 (second column), and altered image (last column).

The ability of an attack to allow a specific intrusion or disruption or cause general chaos is known as adversarial specificity. Based on the attack's specificity, threat models in deep FR systems could be divided into the following categories. A targeted attack causes a model to predict the adversarial example's label incorrectly. This is accomplished by impersonating famous people in a biometric or FR system. As long as the outcomes are not the correct labels, non-targeted attack predicts the labels of the adversarial examples in an irrelevant manner. Face dodging is used to accomplish this in FR/biometric systems. Numerous studies have demonstrated various methods of fooling FR systems. A non-targeted attack is easier to implement than a targeted attack because it has more options and space to alter the output. Additionally, accessories were found to be effective; suggested wearing glasses with adversarial frames made with gradient-based techniques. An improved version of the adversarial eyeglass frames was later made using GAN techniques. To deceive the cutting-edge ArcFaceFR model, printed an adversarial paper sticker and attached it to a hat. These techniques, on the other hand, may make a person stand out from the crowd due to their unnatural appearance when applied to them. In contrast, we propose applying the perturbation to a face mask, a common safety precaution during the COVID-19 outbreak; In addition, our universal attack can be used more broadly without the need for an expert to train a tailor-made one, in contrast to previous work in which the proposed attacks craft tailor-made perturbations (target a single image or person). In addition, we provide a previously unaddressed aspect of our method's efficacy in a real-world use case involving a CCTV system.

2.2 Face Recognition

For identity authentication, face recognition (FR) has been a common biometric technique that is widely used in a number of fields, including finance, the military, public security, and everyday life. The ultimate objective of a typical FR system is to verify or identify a person based on a digital image or video frame taken from a video source. According to the researchers, FR is a biometric application based on artificial intelligence that can only identify a person by analyzing patterns in their facial features. The 1960s provided the inspiration for the concept of using the face as a biometric trait, and the early 1990s saw the creation of the first successful FR system. Recent developments in deep learning, as well as the use of increasing hardware and a lot of data, have led to a significant increase in the number of FR algorithms that perform well. Because of this capability, FR technologies can be widely implemented in a wider variety of applications, such as photo tagging on social media and questionable identification in automated border control systems. However, in an advanced FR model, the system should also be able to withstand potential types of attacks designed to target its proficiency. Promoting recognition efficiency alone is not sufficient.

Researchers recently discovered that FR systems are susceptible to a variety of attacks that create data variations to deceive classifiers. These attacks can be carried out (a) through physical attacks, which alter a face's physical appearance prior to image capture, or (b) through digital attacks, which alter the captured face image. Spoofing attacks, also known as presentation attacks, are frequently employed in physical attacks. Oppositional attacks, morphing-attack variations are essential methods for digital invasion. Keep in mind that although some adversarial attacks are designed to be carried out physically[3].

3. Methods

3.1 Mask Projection

To digitally train the adversarial mask, we need to simulate a mask on a person face in a real world. Hence, It has use 3D face construction to digitally apply a mask on a facial image. an end-to-end approach called UV position map that records the 3D coordinates of a complete facial point cloud using a 2D image. The map record the place information of 3D face and provide dense correspondence to the semantic meaning of each point in UV space, allow us to achieve a near-real approximated of mask on a face, which is a essential to a creation of the adversarial mask in a real world. More formally, then it consider a mask and a rendering function . The rendering function partially inspire from takes a mask and a facial image interface, and applies a mask on the face, result in a masked face image.

Steps of the mask's projection on a facial image is as follows:

1. Detect a landmark points of a face - given the landmark detector, then we extract a landmark points of a face.

2. Map a mask pixel to the facial image - the landmark points of a face extract in the previous step of a pipeline are used to map a mask pixels to a corresponding location on a facial image.

3. Extract depth feature of the face - a facial image is passed to 3D face reconstruction model to obtain depth feature.

4. Transfer 2D facial image to a UV space - the depth features is used as to remap a facial image to a UV space.

5. Transfer 2D mask image to a UV space - the depth feature is used to remap a mask image on a UV space.

6. To improve a robustness of a adversarial mask, random geometric transformations and colour - based augmentations is applied:

(i)geometric transformations - random translation and rotation is added to simulate possible distortions in a mask's placement on the face in real world, and

(ii)colour-based augmentation - random contrast, brightness, and noise are added to simulate changes in the appearance of the mask that might result from various factors.

3.2 Patch Optimization

Image patching provides the capability to choose arbitrarily shaped areas on an image and substitute them with an exterior case to distinct arbitrarily shaped areas, simultaneously with an artificial noise element. This is a conceptual way to remove unwanted faults from an image for ornamental cases. To solve the optimization, there is a two-step approach that involves patch search and vote and reconstruction. Experimental results show that this approach can produce high-quality texture maps better than existing techniques for objects scanned by consumer depth cameras such as Intel RealSense.

The goal of most image-based texture mapping approaches is to produce a high-quality view-independent texture map using a set of N source images, S1, • • •, SN, taken from different viewpoints. A simple way to produce a texture map is to project the source images onto the geometry and combine all the projected images. Ideally, these projected images are photometrically consistent, and thus, combining them produces a high-quality texture map[4,5].

4. Evaluation

We did this research on face recognition in two ways, by considering adversarial effects and without adversarial effects, let us continue with the description, of what we actually need in this project and what kind of recognizers we implemented. We first installed required libraries such as OpenCV, a Computer Vision Library, another library is opencv-contrib which supports additional opencv contribution packages, to the recognizer such as LBPHFace Recognizer.

4.1 About Recognizers

A. FisherFace Recognizer :

Fisherfaces algorithm extracts principle components that separates one individual from another. So, now an individual's features can't dominate another person's features. Fisherface method will be applied to generate a feature vector of facial image data used by the system and then to match the vector of traits of the training image with vector characteristics of the test image using Euclidean distance formula[6].

Fig-2: FisherFace process

B. LBPHFACE Recognizer :

Local Binary Pattern(LBP) is a simple yet very efficient texture operator which labels the pixel of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. It doesn't look at images as a whole but instead tries to find their local structure by comparing each pixel to its neighboring pixels[7].

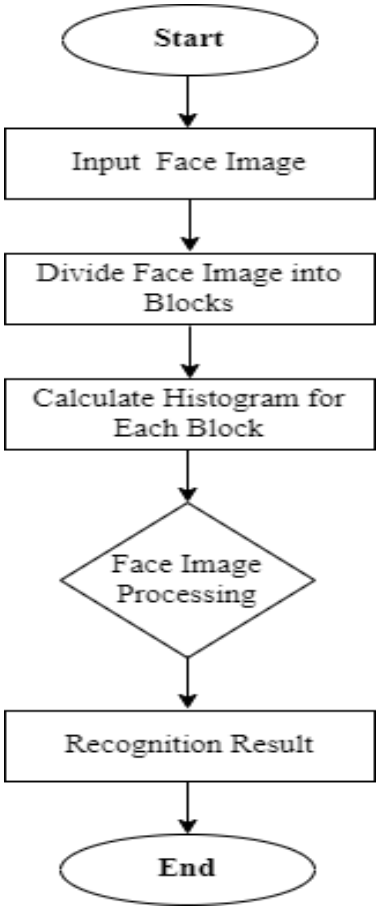

Fig. 3: LBPH Algorithm Flowchart

LBPH uses 4 parameters

a. Radius: To build the circular local binary pattern and represents the radius around the central pixel. It is usually set to 1.
b. Neighbors: The more sample points you include, the higher the fractional cost. It is usually set to 1.
c. X Grid : The number of cells in the horizontal direction.
d. Y Grid : The number of cells in the vertical direction.

These recognizers are helpful in identifying the person as it does classify as zero or one**.**

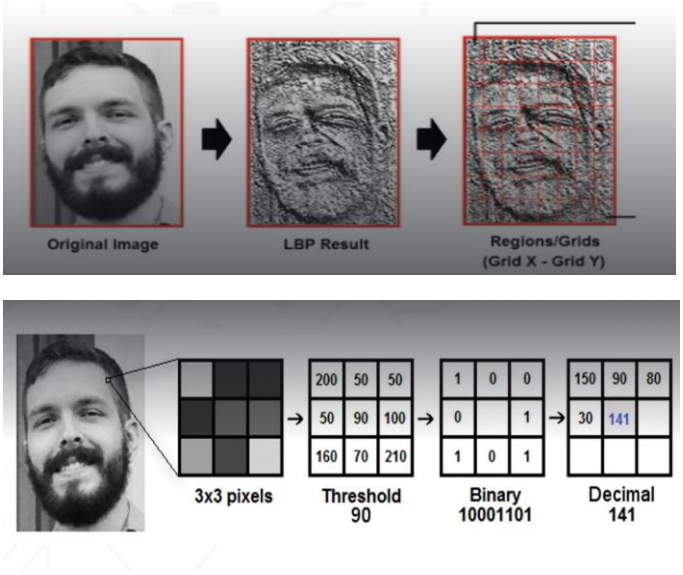

Fig.4: LBPH

Steps we did while predicting are as follow:

a. Loading face detection algorithm that's Harr Cascade Frontal Face algorithm
b. Loading Classifier for Face Recognition
c. Training Classifier for our dataset
d. Reading frame from the camera and pre-processing
e. Face detection by its algorithm
f. Predicting face by loading frame into the model
g. Displays recognized class with its accuracy

We created a dataset by running one python code that captured the faces of people. we captured 20 to 30 images for a person, so for training our model we have a dataset, that is the faces of 2 to 3 people and on that dataset, we trained the model and tested also, and we got an accuracy, of 50% by using LBPHFace Recognizer by considering the adversarial effects and 70% accuracy without considering adversarial effects.

4.1 Digital Attack

Presentation attacks on face recognition systems are categorized into two classes physical and digital. Digital attacks similar to morphing possess entered definite immersion. With the advancements in deep learning and computer vision algorithms, several effortless operations are accessible where with limited gate clicks, an image can be fluently and seamlessly modified. Also, the production of artificial images or updating images, and videos (e.g. creating deep fakes) is somewhat simple and much more efficient due to the tremendous enhancement in generative machine learning models. Numerous of these methodologies can be applied to attack face recognition systems[8].

4.2 Physical Attacks

In this way, to judge the effectivity of attack in the real world, then we print the digital pattern on two surfaces: on regular paper cut in the shape of a face mask and on white fabric mask. Similarly, we create a testbed that operates an end-to-end fully automated FR system simulating a CCTV use case.

The system contains:

(a)Network camera which records a long corridor,

(b)An MTCNN detection model for face detection, pre-processing, and alignment, and

(c)Attacked model - we perform the white-box attack in which a model used for training the adversarial mask is also the model under attack.

Also, we perform a "offline" analysis in black-box setting, on which the facial image is cut from the original frame and compare to ground-truth embedded vectors generated using other models. To calculate the specific verification threshold (set at 0.35), we use a subset of 1,500 identities from the Facial dataset and perform a process. Various face masks are used (digitally) to each original facial images. Then, we calculate the sine similarity between the identity's embedding vector and each masked face image. Since we employ a semi-critical security use case (CCTV), we chose the threshold that led to a false acceptance rate (FAR) of 2%. Furthermore, to minimize false positive alarms, we used the persistence threshold to recognize = 6 frames and a sliding window of sliding window N = 10 frames to designate a candidate identity as the valid one. We consider a group of 10 male and 10 female participants .Each participant has walk along the corridor six times, once with each mask evaluated similar to the digital experiments, and two more times with our adversarial masks printed on paper and fabric. The ground-truth embedding of each participant was calculated using two facial images, where a standard face mask was applied to each image, for a total of four facial images. The results of experiments where we can see that our adversarial masks performed significantly better than the other masks evaluated on every metric, with a high correlation to the cosine similarity results obtained in the digital domain[9].

In original case of CCTV use in which an attackers tries to evaluate a detection of a system, our adversarial fabric mask was able to conceal a identity of 31 out of 35 participants as opposed to a control masks which are able to conceal 8 out of 35 participants at most (persistence detection value is 85.41%). We had also examined the effectiveness of a masks on model as they are not trained. The results show that the masks has similar adversarial effect on FR models in the black-box setting as in the white-box setting. Another aspect are examined in the physical evaluation is the ability to print a adversarial pattern on the real surface. Due to the limited ability the accurate output of the original colours on the fabric, we can recognise that there is a slight difference in a performance of the masks.

5. Countermeasures

We proposed 2 ways 1) Without considering the adversarial attacks that clear face without any masks and we got an accuracy of 70% and 2) By considering an adversarial attack such as masking in a different manner, such as disposable masks(Blue mask) and no disposable (cloths made masks), and that gave us accuracy of 50%.

6. Conclusion

In this study, we suggested a physical global attack against FR systems that takes the form of a face mask. Due to the extensive use of face masks during the COVID-19 epidemic, our mask will not raise any suspicions while other attack tactics used different accessories that are more obvious and do not mix in with the surroundings. We proved the usefulness of our mask both in white-box and black-box scenarios in the digital sphere. In the physical domain, we showed how our mask is able to prevent the detection of multiple participants in a CCTV use case system. Moreover, we proposed possible countermeasures to deal with such attacks. To sum up, in this research, we highlight the potential risk FR models face from an adversary simply wearing a carefully crafted adversarial face mask in the COVID-19 era.

7. References

[1] Submitted on 9 Jun 2022] ReFace: Real-time Adversarial Attacks on Face Recognition Systems Shehzeen Hussain, Todd Huster, Chris Mesterharm, Paarth Neekhara, Kevin An, Malhar Jere, Harshvardhan Sikka, Farinaz Koushanfar

[2]Adversarial Attacks against Face Recognition: A Comprehensive Study. Fatemeh Vakhshiteha, Ahmad Nickabadib, Raghavendra Ramachandra.

[3] Adversarial Mask: Real-World Universal. Adversarial Attack on Face Recognition Models, [0000−0003−0270−1743]

[4] Differentiable Patch Selection for Image Recognition Jean-Baptiste Cordonnier*, Aravindh Mahendran*, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, Thomas Unterthiner CVPR 2021

[5] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. 2017. Patch-Based Optimization for Image-Based Texture Mapping. ACM Trans. Graph. 36, 4, Article 106 (July 2017)

[6] Semi-Automatic Assessment of Modeling Exercises using Supervised Machine Learning. Stephan Krusche, Technical University of Munich.

[7] International Journal of Scientific Research in Computer Science and Engineering, Real Time Face Recognition of Human Faces by using LBPH, Vol.6, Issue.5, pp.06-10, October (2018)

[8] Front. Artif. Intell., 08 December 2021 Sec. Machine Learning and Artificial Intelligence Volume 4 - 2021 | https://doi.org/10.3389/frai.2021.643424

[9] Z. Lim, S.N. Davis, L.S. Goldstein: Making an invisibility cloak: Real world adversarial attacks on object detectors. In: European Conference on Computer Vision. pp. 1–19. Springer (2021)