CSEC-720 Deep Learning Security

Project Proposal

Mehul Sen[1]

[1]Golisano College of Computing and Information Sciences, Rochester Institute of Technology

April 17, 2023

1 Introduction

Facial recognition, a biometric technique, has gained popularity recently due to rapid advancements in AI and machine learning technologies. This progress has led to the development of high-accuracy Face Recognition Systems (FRS), offering a passive and non-intrusive means of verifying individual identities.

Despite the advantages of facial recognition technology, concerns exist about its potential misuse, particularly with privacy and security. FRS is particularly vulnerable to adversarial attacks, as first proposed by Szegedy et al. [21]. Sharif et al. discussed how these attacks could result in misclassifications and enable attackers to manipulate the system to falsely incriminate innocent individuals [18]. Consequently, the risk of such attacks is a serious concern and has driven additional research in this area. The objective is to develop new defenses and enhance existing ones, thereby making facial recognition systems more resistant to adversarial attacks.

In my literature review, I analyzed numerous pivotal papers covering face recognition systems (FRS), adversarial attacks on image classification models, and studies addressing adversarial attacks on FRS. My research indicated that numerous face recognition models are trained on datasets with limited variations within their poses, facial expressions, illuminations, etc. [5, 8, 10], which leaves them susceptible to adversarial attacks.

General image classifiers are trained for broader object recognition tasks and focus on various objects or scenes, capturing features such as edges, shapes, textures, and colors. In contrast, face recognition systems are specifically tailored for identifying and verifying individual human faces [8]. These systems concentrate on learning specific and detailed features, such as the shape of the eyes, nose, mouth, facial landmarks, and the spatial relationships between these facial components.

This project aims to evaluate the performance of adversarial attacks on FRS, specifically the Customized Iteration and Sampling Attack (CISA). Traditional decision-based attacks perform better in tasks involving strong prior knowledge, such as face recognition [19]. CISA takes this further by utilizing transfer-based attacks to generate an intermediate adversarial example as the starting point for its decision-based attack implementation. CISA is designed to concentrate on the most influential pixels. Through a customized iterative process, it exploits the sensitivity to pixel perturbations within facial components that FRSs heavily rely upon. By combining these two attack strategies and implementing a black-box design, CISA would make an effective attack against FRS.

Utilizing the DeepFace model trained on the UTKFace datasets, this project will offer a specialized analysis of adversarial attacks in facial recognition. The aim is to identify a computationally inexpensive black-box adversarial attack that does not require additional accessories and is not easily detectable. By testing the effectiveness of the CISA attack on FRS and comparing its performance with other adversarial attacks, the project will contribute to a better understanding of the effectiveness and generalizability of these attacks on various face recognition models. Ultimately, the findings could inspire future research on developing robust defenses against adversarial attacks, leading to more secure and reliable face recognition systems.

2 ML/DL Background

Neural networks have been widely used for Image Classification Tasks - categorizing images into specific classes based on visual content [8]. A common application is in Facial Detection and Recognition Systems. Face Detection involves finding a face in an image and processing it for easier recognition. FRSs then compare this processed face to a database of known faces to identify the individual.

There are two categories of FRSs [10] : one identifies a person within a large database of faces and returns a list of likely matches; this system typically has few available images per person, and real-time recognition is unnecessary. The second category involves identifying people in real-time through security monitoring; multiple images are available for training, and real-time recognition is required.

Some popular algorithms used in FRS include Local Binary Pattern (LBP) [1], Eigenfaces and Principal Component Analysis (PCA) [10], Linear Discriminant Analysis (LDA), and the PAL Face Recognition Algorithm.[5]

A significant limitation of several proposed FRS models is their need for more variation in lighting conditions, backgrounds, expressions, and occlusions. This vulnerability to adversarial attacks highlights the need for further research to implement defenses against such attacks on FRS.

3 Security Background and Threat Model

Security Background. Neural networks are vulnerable to adversarial attacks[21]. One example is in FRS as image classifiers, where an adversarial attack can fool the system into misclassifying a face[15].

Some common techniques for carrying out attacks against FRS include:

- Adding noise or obstructions: This can be done by using additional accessories such as masks, glasses, or makeup to disrupt the features used by face recognition models to identify faces. [11]
- Adversarial examples: These examples can be applied to a person's face, causing the model to misclassify their identity. [15, 18]
- Warping and Masking: This technique involves using face warping or combining two or more images of different individuals to impersonate multiple individuals. This can contour the features an FRS might rely on to identify an individual. [9]

These attacks are designed to deceive a classifier into misclassifying a specific input in a dataset. However, universal perturbations are adversarial perturbations that can be added to any input in a dataset, causing the neural network to misclassify it. These perturbations are designed to cause misclassification across the entire dataset and require minimal computation during the attack. [17]

Threat Model. Most of the papers discussing adversarial attacks have one of the following three primary goals for attackers:

- Deceiving the FRS into not identifying the face from any image, potentially allowing an attacker to bypass overarching crime detection and prevention systems.
- Dodging attacks, also known as untargeted attacks, where an attacker aims to have their face misidentified as any other arbitrary face. This type of attack leads to misclassification of the input. These attacks are not necessarily malicious and can also be used by individuals seeking to protect their privacy against excessive surveillance. [18, 23]
- Impersonation attacks, also known as targeted attacks, where an adversary attempts to recognize their face as another. This type of attack changes the output classification of the input to a desired one. For example, an attacker could attempt to confuse law enforcement by simultaneously tricking multiple geographically distant surveillance systems into detecting their presence in different locations or manipulating evidence and framing someone for a crime. [18, 23]

These papers assume the attacker can access the FRS and intends to launch a dodging or impersonation attack against an already trained system. The attackers cannot tamper with the face recognition model's training data by injecting mislabeled or altered training data. Instead, their actions are limited to modifying the model's input data composition for classification based on their understanding of the classification model.

Many papers employ a white-box approach and are familiar with the internal workings of the model. However, papers such as [18, 16, 7, 6] take a black-box approach, where the attackers lack detailed knowledge about the internal workings of the system.

In some attacks, attackers go beyond modifying the input data and set up additional components to aid their efforts. For instance, in [15], attackers set up additional facial photos with a chessboard pattern to improve their attack. In [11], they created a makeup dataset to help them evade the system's facial recognition. Other attacks, such as [18, 23], involve designing adversarial facial accessories that can distort or obfuscate the target's facial features and deceive the system.

For this paper, I will adopt a black-box approach in which the attacker cannot directly access the FRS and has a limited number of queries. I will assume that the attacker knows enough about the target model to develop their substitute model. I aim to perform a dodging attack to misclassify faces as any other face within the training dataset. This approach simulates the most likely scenario for FRS, yielding realistic results.

4 Related Work

Adversarial Attacks on Image Classifiers. Adversarial attacks have become a prominent area of research due to the widespread use of machine-learning models as image classifiers in various applications. Recently, plenty of innovations have been in generating adversarial examples to misclassify image classifiers.

Shafahi et al. [17] propose a universal adversarial attack against image classifiers by going through several iterations of updating the neural network weights using gradient descent and then updating the universal perturbations using ascent. The authors also propose a defense mechanism that involves training models on these perturbations to improve the robustness of the model. However, a limitation of [17] is that it may result in an imbalanced level of robustness across the classifier's classes making the attack relatively ineffective against all classes, unlike targeted adversarial attacks that can be tailored to specific classes.

Rahmati et al. [16] introduce a new framework called Geometric Decision-based Attack (GeoDA) for generating adversarial examples against image classifiers in a black-box setting where only the top-1 label of the classifier is available. The framework leverages the fact that the decision boundaries of the model have a small mean curvature near data samples. This property can be used to design query-efficient attacks. The authors demonstrate that GeoDA outperforms other state-of-the-art black-box attacks. Although GeoDA can efficiently perform black-box attacks, it may not withstand countermeasures such as gradient masking [13] or randomization [22]. These measures can modify the decision boundary that GeoDA heavily depends on to produce its adversarial examples.

Zhou et al. [24] review recent literature on adversarial attacks and their countermeasures in deep learning. They discuss traditional attacks, such as optimization-based, gradient-based, and GAN-based attacks, which require full knowledge of the target models and are primarily white-box attacks. In contrast, advanced attacks like universal adversarial, transfer-based, and query-based attacks face two main challenges. Transfer-based attacks often have low success rates due to insufficient adjustment procedures for information from surrogate models. Meanwhile, query-based attacks can result in an extremely high number of queries.

Shi et al. [19] developed the Customized Iteration and Sampling Attack (CISA), a query-efficient black-box adversarial attack method that combines transfer-based and decision-based attacks. In contrast to traditional gradient-based attacks, CISA utilizes a transfer-based component to generate an intermediate adversarial example, which is compressed and refined through a decision-based attack process. This approach yields a query-efficient adversarial attack with significantly reduced noise magnitude and superior performance compared to other state-of-the-art methods.

Adversarial Attacks on Face Recognition Systems. FRS are among the most commonly used applications of image classifiers; research has been conducted to study adversarial attacks on these systems.

Sharif et al. [18] focused on making their attacks inconspicuous and physically realizable. They attacked FRSs based on NNs, using three NNs based on a design by Parkhi et al. [14] for a white-box attack. They attacked it with eyeglasses imprinted with texture to perform impersonation and dodging. Using a query-based attack, they also conducted a black-box attack on the commercial FRS

Face++. They achieved an almost complete success rate and proved that while humans are unaffected by small changes to images, machines can be greatly affected.

Pautov et al. [15] used a more practical approach by simulating a real-world adversarial attack on FRSs. They applied adversarial patches to a person's clothing or face, which were then used to fool LResNet100E-IR with ArcFace loss. They successfully fooled the FRS and discovered that the bigger the patch and the closer it is to the eyes, the better the attack results.

Lin et al. [11] attempted to fool the VGG16 classifier using a Cycle-GAN to generate adversarial makeup. Their attack showed that makeup could alter facial features and create perturbations that successfully mislead FRS. Their method worked on an untargeted model with pre-trained weights and a targeted model trained from scratch. They collected their image dataset to evaluate their attack and created their face recognition model.

Zheng et al. [23] proposed a novel robust physical attack framework called PadvFace that specifically considers and models physical-world condition variations. It is designed to study the sticker-based physical attacks that aim to generate wearable adversarial stickers to deceive state-of-the-art face recognition. PadvFace can generate adversarial stickers more robust to physical-world conditions, such as lighting changes and camera angles. It can also generate adversarial stickers that are more difficult to detect by the human eye. They tested and discovered that their attack was effective on both Dodging and Impersonation attacks. One of the major drawbacks of PadvFace is that it is not inconspicuous in the physical world, and the adversarial perturbations are noticeable.

The studies conducted in [11, 15, 18] focused on limited FRSs and datasets. However, including a broader range of FRSs and datasets could provide further insights into what works and what does not. It is worth noting that papers [11, 15, 23] are resource-intensive and necessitate significant effort and resources, such as additional accessories to be used before a successful attack can be carried out. Research has also explored attacks designed to digitally perturb images without involving any extra accessories, makeup, or adversarial patches.

Cherepanova et al. [3] developed a new evasion tool called LowKey that generates perturbed images from original images. When FRS uses these modified images, the systems fail to predict subsequent images of the subjects protected with LowKey accurately. The authors demonstrate that LowKey is highly effective against commercial black-box APIs such as Amazon Rekognition and Microsoft Azure Face. Additionally, LowKey is scalable, robust to image compression, computationally fast, and transferable to other models. A major drawback of LowKey is that it requires the face recognition model to be trained using images that are already protected. Therefore, if the model has already been trained using original images, this tool might not work.

Kasichainula, Mansourifar, and Shi [9] introduce a new adversarial attack on FRS called Recursive Adversarial Attack (RAF). RAF employs smart face warping techniques such as increasing the smile, raising eyebrows, and stretching the nose. It also uses a depth-first search tree with recursive optimization to decompose high-dimensional optimization problems into more manageable sub-problems. The authors demonstrate that their black-box attack setting can fool many FRS models and find an adversarial instance within just six queries. The RAF attack assumes that the black-box face recognition model allows attackers to conduct at least three queries without taking defensive measures toward the query attack.

Furthermore, neither the attacks proposed in [3] nor [9] have been tested against adversarially trained defenses, which incorporate samples of adversarial attacks as part of their defense mechanisms. As a result, these attacks may not be as effective in such scenarios. Moreover, these attacks might be easy to detect due to the unnatural modifications to facial features or the gradient-based perturbations added to the images. This has also motivated research into implementing adversarial images that are more difficult to detect and conform to facial features rather than completely altering them.

Jia et al. [7] proposed a novel black-box adversarial face-recognition attack method called Adv-Attribute. This method crafts the adversarial noise and injects it into multiple attributes based on the guidance of the difference in face recognition features from the target. Unlike gradient-based or patch-based attacks, this attack appears more natural and invisible. It uses StyleGAN for face generation and a flexible multi-objective optimization paradigm better to balance the trade-off between stealthiness and attacking strength. Adv-Attribute was tested on IR152, Facenet, and Mobileface models with datasets CelebA-HQ for face images and FFHQ to train the GAN model.

Hu et al. [6] proposed a new framework to generate adversarial face images with a natural appearance and strong black-box attack strength called Adversarial Makeup Transfer GAN (AMT-GAN).

AMT-GAN can attack various FRS by using adversarial makeup transfer to transfer makeup from reference images to adversarial images, making them more natural and comfortable. AMT-GAN was tested on IRSE50, IR152, Facenet, and Mobileface models with datasets CelebA-HQ for face images and LADN for makeup. As mentioned by the authors, AMT-GAN has a higher attack strength and better visual quality in images of females caused by the unbalance of gender in the makeup transfer training dataset.

Another limitation of [7, 6] is that since they both use GANs trained on a dataset of real images, the training process of their models is computationally expensive, requiring a large number of iterations of the algorithm. Additionally, some of the datasets used do not have sufficient variations. For example, the CelebA-HQ dataset has significant background clutter and some pose variation, but it does not consider lighting conditions.

Based on the papers mentioned above, it is evident that there is a need for computationally inexpensive black-box adversarial attacks that do not require any additional accessories and are not easily detectable. This should be achieved while maintaining a high success rate and query efficiency. Furthermore, the attack should be tested on models specifically designed for face recognition, trained with datasets encompassing a variety of expressions, poses, and lighting conditions.

5 Project Idea

In this project, I will conduct an experimental evaluation of CISA and compare its performance against two popular gradient-based adversarial attack methods: Projected Gradient Descent (PGD) [12] and Evolutionary Attack (EVO) [4]. CISA combines both transfer-based and decision-based adversarial attacks. We need to evaluate both the transfer and decision-based components to assess their effectiveness.

PGD can maintain a tolerable visual appearance even with an increased epsilon value [20]. It will be implemented as a comparative transfer-based attack, generating adversarial examples on a substitute model, which are then used to attack the target model.

On the other hand, EVO can constrain adversarial noise to the central part of images and leverage prior knowledge during attacks, making it effective against FRS [19]. It will be implemented as a comparative decision-based attack to further improve upon adversarial noise from the PGD attack.

To develop a more accurate face recognition system (FRS) model that better reflects real-world scenarios, my project will employ the DeepFace model, a lightweight face recognition and facial attribute analysis framework, and a hybrid of state-of-the-art face recognition models found on GitHub. This would allow us to create an FRS model that is more accurate and representative of real-world situations allowing us to more effectively test the efficacy of CISA and other adversarial attacks on face recognition systems. This model will be trained on the UTKFace dataset, which consists of over 20,000 face images with annotations of age, gender, and ethnicity. It covers a wide range of variations in pose, facial expression, illumination, occlusion, and resolution and would the robustness of the attack with variations in image conditions. The transfer-based attack component requires a substitute model to generate the initial adversarial examples. I will implement the VGGFace2 model [2], an enhanced version of the VGGFace model employed in DeepFace as a white-box substitute attack model, which the attacker will use to develop their intermediate adversarial examples.

My project will focus on the performance of adversarial attacks on FRS rather than generic image classifiers. This project will build upon existing research that employs models trained on datasets containing various object classes, which are then adapted into face recognition models trained to emphasize the facial components within an image. This project aims to enhance researchers' understanding of the effectiveness of black-box adversarial attacks in the context of FRS and to shed light on the generalizability of these attacks across different face recognition models. This insight could inspire future research on developing defenses against adversarial attacks, ultimately leading to more secure and reliable face recognition systems.

The estimated timeline of my project is as follows:

- Week 1(3/30 - 4/5): Implement the DeepFace model and evaluate its performance on the UTKFace dataset alongside the VGGFace2 [2] substitute model. (Estimated time: 10 hours)
- Week 2(4/6 - 4/12): Recreate CISA as described in [19] and replicate their findings. (Estimated time: 10 hours)

- Week 3(4/13 - 4/19): Employ the CISA attack on the DeepFace model and evaluate its effectiveness. Additional time will be used for testing and debugging. (Estimated time: 10 hours)
- Week 4(4/20 - 4/26): Employ the PGD and EVO attacks on the DeepFace model and evaluate its effectiveness. (Estimated time: 6 hours)
- Week 5(4/27 - 5/3): Aggregate the results and compile a comprehensive report listing all the findings. (Estimated time: 4 hours)

During this project, I may face challenges in replicating the results of [19] and ensuring a fair comparison between attacks. To tackle these obstacles, I will engage in careful planning, break tasks into manageable sub-tasks, and prioritize crucial objectives facilitating a smoother execution of the project.

During Weeks 4 and 5, my class schedule will be quite heavy, which is why most of the progress will be made during the initial weeks of the project.

6 Appendix

ChatGPT, developed by OpenAI, was utilized to brainstorm and enhance the quality of the paper.

Prompt Used: 'Improve the grammar and flow of this paragraph'

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. "Face Description with Local Binary Patterns: Application to Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2006).

[2] Qiong Cao et al. "VGGFace2: A Dataset for Recognising Faces across Pose and Age". In: *IEEE International Conference on Automatic Face & Gesture Recognition* (2017).

[3] Valeriia Cherepanova et al. "LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition". In: *International Conference on Learning Representations* (2021).

[4] Yinpeng Dong et al. "Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019).

[5] Mahmood Ul Haq et al. "COMSATS Face: A Dataset of Face Images with Pose Variations, Its Design, and Aspects". In: *Mathematical Problems in Engineering* (2022).

[6] Shengshan Hu et al. "Protecting Facial Privacy: Generating Adversarial Identity Masks via Style-robust Makeup Transfer". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[7] Shuai Jia et al. "Adv-Attribute: Inconspicuous and Transferable Adversarial Attack on Face Recognition". In: *Conference on Neural Information Processing Systems* (2022).

[8] Patrik Kamencay et al. "A new method for face recognition using convolutional neural network". In: *Advances in Electrical and Electronic Engineering* (2017).

[9] Keshav Kasichainula, Hadi Mansourifar, and W. Shi. "RAF: Recursive Adversarial Attacks on Face Recognition Using Extremely Limited Queries". In: *IEEE International Conference on Big Data* (2022).

[10] Steve Lawrence et al. "Face Recognition: A Convolutional Neural-Network Approach". In: *IEEE transactions on neural networks* (1997).

[11] Chang-Sheng Lin et al. "Real-World Adversarial Examples Via Makeup". In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2022).

[12] Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *ArXiv* (2017).

[13] Nicolas Papernot et al. "Practical Black-Box Attacks against Machine Learning". In: *Asia Conference on Computer and Communications Security* (2017).

[14] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: *British Machine Vision Conference* (2015).

[15] Mikhail Aleksandrovich Pautov et al. "On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System". In: *International Multi-Conference on Engineering, Computer and Information Sciences* (2019).

[16] Ali Rahmati et al. "GeoDA: A Geometric Framework for Black-Box Adversarial Attacks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020).

[17] Ali Shafahi et al. "Universal Adversarial Training". In: *AAAI-20 Conference on Artificial Intelligence* (2018).

[18] Mahmood Sharif et al. "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition". In: *SIGSAC Conference on Computer and Communications Security* (2016).

[19] Yucheng Shi et al. "Query-Efficient Black-Box Adversarial Attack With Customized Iteration and Sampling". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[20] Sameer Singh. "Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks". In: *Handbook of Digital Face Manipulation and Detection* (2022).

[21] Christian Szegedy et al. "Intriguing properties of neural networks". In: *Computing Research Repository* (2013).

[22] Cihang Xie et al. "Mitigating Adversarial Effects Through Randomization". In: *International Conference on Learning Representations* (2018).

[23] Xin Zheng et al. "Robust Physical-World Attacks on Face Recognition". In: *Pattern Recognition* (2021).

[24] Shuai Zhou et al. "Adversarial Attacks and Defenses in Deep Learning: From a Perspective of Cybersecurity". In: *ACM Computing Surveys* (2022).