Adversarial Attacks on Face Recognition Systems

1st Shambhavi Jha
220953634
Department of ICT
Manipal Institute of Technology
Email: shambhavi.mitmpl2022@learner.manipal.edu

2nd Archit Singh
220911464
Department of ICT
Manipal Institute of Technology
Email: archit2.mitmpl2022@learner.manipal.edu

3rd Lakshay Prasher
220911011
Department of ICT
Manipal Institute of Technology
Email: lakshya.mitmpl2022@learner. manipal.edu

4th Krish A Manchanda
220953660
Department of ICT
Manipal Institute of Technology
Email: krrish.mitmpl2022@learner. manipal.edu

***Abstract*—Systems for face recognition have diverse applications, ranging from security to surveillance to identity verification. Such systems, however, can be attacked by opponents who may add small, imperceptible perturbations to the images in order to mislead the AI models. The project tackles adversarial attacks on face recognition systems using the Fast Gradient Sign Method (FGSM) and DeepFool algorithms. The attack implementation within this study follows an end-to-end face recognition pipeline, which includes MTCNN for face detection and FaceNet for feature extraction and classification. The facial imagery for this study is downloaded from various public sources, including CelebA. The preprocessing of the images practically includes some combination of resizing, normalization, and face alignment, which were somehow deemed important for the model performance. The main thrust of this project is to show how adversarially generated images of faces can be effectively used against recognition models in turn to generate the misclassification of identities. Hence, arising issues pose severe threats to the safety of possible systems, along with the need for address against them in face recognition.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

The adoption of face recognition systems in security and authentication applications and even surveillance applications has been great. Deep learning models are used in these systems to identify individuals adequately by facial features. Despite their effectiveness, they are still susceptible to a class of attacks referred to as adversarial attacks-in which a small, carefully crafted perturbation in the image, misleads the model to classify it incorrectly. Such attacks can prove very severe security risks, especially in high-stakes applications like biometric authentication, law enforcement, and in financial transactions. The project examines how adversarial attacks can affect a face recognition system: by generating adversarial faces using techniques like Fast Gradient Sign Method and DeepFool.

Achieve this with a face recognition pipeline MTCNN face detection and FaceNet feature extraction/classification. Add adversarial perturbations to the faces in the images, and study the effects on model performance. Appropriate preprocessing steps such as image resizing, normalization, and face alignment should be included to improve accuracy. The aim of this project is to demonstrate that adversarial images are indeed capable of misleading AI-based facial recognition systems and to strengthen the argument for adversarial defenses in real-life applications.

The present paper is divided into seven sections: A literature survey, Section 2, reviews previous works in the adversarial field of face recognition, analyzing attack methodologies, and their effect on many applications using deep learning. Section 3: Works objectives outline the principal objectives of the project and provide an overview of examing each adversarial attack in face recognition. Section 4: Dataset illustrates this dataset, from CelebA, with preprocessing techniques such as resizing, normalization, and alignment. The basis of operation is presented in Section 5: Methodology, in a stepwise implementation of the whole project including detection and recognition of faces, adversarial attack generation using FGSM and DeepFool, and evaluation of the performance of the model under attack. Section 6 deals with findings, exposes vulnerabilities detected in face recognition models, and proposes defenses. Finally, Section 7 lists all research papers, books, and other internet sources consulted for this study.

II. Literature Survey

Facial recognition technology is now an inherent component in multiple security and verification contexts. Since it is susceptible to adversarial attacks—purposeful manipulations meant to trick algorithms—it remains a serious issue in terms of security. Data preprocessing has stepped up as the essential approach that helps strengthen the technology against attacks. This work is a bibliometric analysis that explores current work published in IEEE Xplore focusing on facial recognition adversarial attacks and the importance of preprocessing in counteracting such attacks.

A. Adversarial Attack Mitigation Strategies Using Pixel Preprocessing and Image Labeling Techniques

This paper introduces a defense strategy that utilizes adaptive preprocessing methods on pixel images and labels to safeguard against adversarial attacks. The method adapts images according to label loss without separating clean and adversarial examples to improve facial recognition system robustness. [1]

B. Measuring the Performance of Efficient Face Anti-Spoofing Detection

The authors explain how various preprocessing and training methods affect the performance of light-weight convolutional neural networks (CNNs) in the task of face anti-spoofing detection. Experimentation using MobileNetV3 shows that the right preprocessing methods can significantly improve the ability of the system to detect adversarial spoofing attacks. [2]

C. An In-Depth Risk Analysis Approach to Adversarial Attacks on Biometric Authentication Systems

This paper gives a risk analysis framework for assessing adversarial attacks against biometric systems, including face recognition. Using scenario-based analysis, the research identifies the vulnerabilities of such systems and stresses the need for strong preprocessing methods to counter the potential risks. [3]

D. Deep Learning and Features Fusion-Based Detection of Adversarial Attacks

The authors present a detection mechanism that combines deep learning methods with feature fusion techniques for adversarial attack identification. During the testing stage, raw data is handled prior to its input into an adversarial discriminator module, thereby enhancing the system's performance in detecting and canceling out adversarial inputs effectively. [4]

E. A New Preprocessing Technique for Reducing Adversarial Examples in Face Recognition Systems

This work presents a new preprocessing method that uses transformation-based defense strategies, e.g., wavelet domain filtering and image reconstruction, to counter adversarial attacks. The work shows that such preprocessing methods are capable of efficiently restoring facial images prior to classification, enhancing the robustness of the model. [5]

III. Objectives

The study's focus is on detecting and exposing the weaknesses of face recognition systems in the face of different kinds of attacks, especially adversarial attacks. A face recognition pipeline is to be built, with FGSM and DeepFool as application-based attack approaches to assess how perturbations tend to affect computer vision and thereby eventually infer possible countermeasures.

- Develop a thorough knowledge of attacks against face recognition systems. Study tiny, barely noticeable perturbations applied to facial images that would deceive deep-learning models into misclassifying images, sometimes leading to severe misclassification.
- Develop a face recognition pipeline- MTCNN can be used for the finding of faces, while FaceNet can be used for feature extraction and classification in order to establish a robust recognition system.
- Attack generation: Generate adversarial faces using adversarial attacks such as FGSM and DeepFool to generate adversarial images used to manipulate the face recognizer.
- To enhance quality and preprocessing of dataset – Preprocessing would involve resizing images, normalizing, and face aligning to make a more accurate and robust recognition model.
- To analyze the impact of adversarial attacks - Estimate the accuracy drops and misclassification rates to assess the effects of adversarial perturbations on model performance.
- To link security vulnerabilities with possible mitigations - Illustrate the real-world threats that adversarial attacks pose and identify possible mitigating measures that could enhance the resilience of a face recognition system.

IV. Dataset

We are using CelebA (CelebFaces Attributes Dataset) for this project, one of the most extensive face datasets, which has gained prominence in research tasks involving computer vision, facial recognition, and adversarial research. The dataset contains more than 200,000 face pictures belonging to 10,177 celebrities and annotated as regards 40 facial attributes such as gender, age, facial expressions, and accessories (like glasses, hats). These were collected from various online sources to assure diversity in illumination, poses, backgrounds, and expressions, making them perfectly fit the robustness testing of face recognition models.
CelebA is selected for this project primarily because it boasts such a rich collection of high-resolution facial images, with major variations in pose, background, and lighting conditions, which would allow us to study the effect of adversarial attacks on face-recognition models under realistic settings, and also secondarily, because it contains aligned cropped facial images that may be straightened to simplify face detection and feature extraction.

V. Methodology

The project the authors followed has a methodology aimed at analyzing and successfully executing adversarial attacks upon a face recognition system using facial detection through MTCNN and feature extraction and classification with FaceNet. Two attack techniques were implemented here-FGSM and DeepFool-for the generation of adversarial images of faces. The attacks were then analyzed for their effectiveness in deceiving the model. The dataset used was CelebA, which is a large-scale face dataset, and the purpose of the preprocessing was to enhance the accuracy and robustness of the model, that is, image resizing, image normalization, and face alignment. Then, the impact of the adversarial attacks was evaluated in terms of the misclassification rate and accuracy of the model,

thereby demonstrating a reduction in accuracy by adversarial attacks.

A. Face Recognition System Setup

- Put in place a face recognition system using MTCNN for face detection, and FaceNet for feature extraction.
- Loading and processing the CelebA dataset such that it leads to highly qualitative face images for the purpose of training and evaluating the models
- Putting into place either training of the models or using a pre-trained FaceNet model to classify faces through their corresponding embeddings.

B. Dataset and Preprocessing

- Used Dataset: CelebA: Contains more than 200,000 celebrity face images tagged with 40 attributes per image.
- Pre-processing Steps:
  - Face Detection: Cutting out the concerned faces from images using MTCNN.
  - Image Resizing: Standardized all images to 160×160 pixel size.
  - Normalization: Scale pixel values between 0 and 1 succeedingly–improves model performance.
  - Face Alignment: The alignment of face landmarks besides being used for face input into FaceNet is because of the same orientation.

C. Generating Adversarial Attacks

- Fast Gradient Sign Method (FGSM):
  - Adversarial images are generated by introducing a small perturbation in the direction of the gradient.
  - Formulated as:$x_{\text{adv}} = x+\epsilon\cdot\text{sign}(\nabla_x J(\theta, x, y))$,where ϵ controls the perturbation strength.
  - Evaluate how varying epsilon values affect the attack's success.
- DeepFool Attack
  - In an iterative process, finds the least perturbation that moves the image across the decision boundary.
  - The perturbation on the adversarial image should not render it visually distinguishable from the original image.

D. Evaluating Attack Effectiveness

- Evaluate the face recognition accuracy prior to and following the application of adversarial attack methods.
- Investigate the misclassification rates, and ex-amine the effectiveness of the deception in FaceNet harm to the model.
- Try different attack strengths to analyze misclassification effects on a model.

VI. Conclusion

This project successfully demonstrates the vulnerability of face recognition systems to adversarial attacks. By implementing a face recognition pipeline using MTCNN for face detection and FaceNet for feature extraction, we evaluated the impact of adversarial perturbations generated using Fast Gradient Sign Method (FGSM) and DeepFool. The experiments on the CelebA dataset revealed that even small, imperceptible modifications to input images can mislead the model into incorrect classifications, highlighting significant security concerns in biometric authentication systems.

The findings emphasize the need for robust defense mechanisms to counter adversarial threats in real-world applications. Techniques such as adversarial training, input preprocessing, and defensive distillation could enhance model resilience against such attacks. This study provides a foundation for further research into improving the security of face recognition models, ensuring their reliability in critical applications such as surveillance, identity verification, and access control systems.

References

[1] Mengqian Li and Chunjie Cao. Defense against adversarial attacks using image label and pixel guided sparse denoiser. In *2022 7th International Conference on Big Data Analytics (ICBDA)*, pages 253–258, 2022.

[2] Luis S. Luevano, Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Miguel González-Mendoza, and Davide Frey. Assessing the performance of efficient face anti-spoofing detection against physical and digital presentation attacks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1021–1028, 2024.

[3] Seong Hee Park, Soo-Hyun Lee, Min Young Lim, Pyo Min Hong, and Youn Kyu Lee. A comprehensive risk analysis method for adversarial attacks on biometric authentication systems. *IEEE Access*, 12:116693–116710, 2024.

[4] Ángel Luis Perales Gómez, Lorenzo Fernández Maimó, Alberto Huertas Celdrán, and Félix J. García Clemente. Detection of adversarial attacks using deep learning and features extracted from interpretability methods in industrial scenarios. *IEEE Access*, 13:2705–2722, 2025.

[5] Alicia Bernice, Kinanthi Marew, and Siti Elda Hiererra. Critical factors that influence the usage of robo advisor application based in indonesia. In *2024 International Conference on ICT for Smart Society (ICISS)*, pages 1–8, 2024.