Convolutional Autoencoder-Based Adversarial Defense for Robust Face Recognition under FGSM White-Box Attacks

[1] Sania Naveed, [2] Zunaira Rafaqat
[1] Chenab Institute of Information Technology, Pakistan
[2] Chenab Institute of Information Technology, Pakistan
Corresponding E-mail: sania.naveed@cgc.edu.pk

Abstract

Face recognition systems have achieved exceptional performance in security, authentication, and surveillance applications. However, these systems are highly vulnerable to adversarial attacks, especially the Fast Gradient Sign Method (FGSM), which perturbs input images subtly but effectively enough to mislead deep learning models. In this paper, we propose a convolutional autoencoder-based defense mechanism to enhance the robustness of face recognition models against FGSM white-box attacks. The approach leverages the capacity of convolutional autoencoders to learn compact and noise-invariant latent representations that can reconstruct clean versions of adversarially perturbed faces. We evaluate the proposed defense on standard face datasets under varying attack intensities and demonstrate significant improvements in recognition accuracy after reconstruction. Our results highlight the potential of autoencoders as a lightweight and effective pre-processing step to mitigate adversarial threats without needing to modify or retrain the target recognition model.

Keywords: Adversarial Defense, Convolutional Autoencoder, FGSM Attack, Face Recognition, White-Box Attack, Deep Learning Robustness

I. Introduction

Face recognition systems have become a core component in identity verification and surveillance technologies due to advancements in deep learning and computer vision. However, their vulnerability to adversarial examples poses a serious security threat. White-box attacks such as FGSM allow attackers to exploit model gradients to generate inputs that appear visually unchanged to humans but cause misclassification by neural networks [1]. This raises urgent questions about the reliability of face recognition models in real-world

applications where robustness to adversarial perturbations is crucial. In this context, it is essential to explore mechanisms that can restore the integrity of facial inputs before recognition. The FGSM white-box attack uses a known model's gradient information to craft adversarial noise, leading to high attack success rates with minimal perturbations [2]. These attacks exploit the linearity and high-dimensional nature of neural networks. Traditional methods to defend against FGSM include adversarial training, input transformation, and gradient masking. However, these approaches either incur high computational costs or degrade model performance on clean data. Moreover, some defenses fail when facing adaptive attacks or unknown attack parameters, making them unreliable in dynamic environments [3].

Convolutional autoencoders (CAEs) have emerged as a promising approach for image denoising, compression, and reconstruction tasks. Their ability to extract semantically meaningful features and ignore irrelevant noise makes them suitable for filtering out adversarial perturbations. By training the CAE on clean face images, the network learns to focus on structural and identity-related patterns while disregarding pixel-level noise introduced by adversaries [4]. This property can be exploited to reconstruct clean images from their adversarial counterparts before they are fed into face recognition systems. This paper proposes a novel adversarial defense mechanism using a convolutional autoencoder to purify FGSM-attacked face images. The defense works as a pre-processing step that does not require access to the internal parameters of the face recognition model. This modularity makes the approach suitable for integration with existing systems [5]. We perform extensive experiments to measure the performance of face recognition models on adversarially perturbed and autoencoder-reconstructed images, thereby quantifying the effectiveness of our method under various attack strengths. The structure of the paper is as follows: we first detail the architecture and training process of the convolutional autoencoder used in our defense. Next, we describe the experimental setup, including datasets, face recognition models, and attack parameters [6]. We then present quantitative results and visual comparisons to demonstrate the efficacy of our approach. Finally, we discuss the strengths and limitations of our method and suggest future directions for robust adversarial defense in face recognition systems.

II. Related Work

The problem of adversarial vulnerability in deep neural networks has attracted growing attention in recent years. Goodfellow et al. introduced FGSM as a method to exploit model gradients, demonstrating that even slight pixel-level modifications can dramatically reduce classification accuracy [7]. Numerous studies have since explored adversarial attack strategies and corresponding defenses in both image classification and biometric domains. Among the most common white-box attacks, FGSM remains a widely studied benchmark due to its simplicity and high efficacy. Defense mechanisms against FGSM attacks include adversarial training, gradient obfuscation, and input preprocessing, each with its benefits and trade-offs. Adversarial training involves incorporating adversarial examples into the model's training data to improve resilience [8]. While effective, this approach is computationally intensive and often reduces accuracy on clean samples. Another category of defenses includes input transformations such as JPEG compression, image quilting, and random resizing, which aim to destroy the adversarial perturbations. However, these transformations may distort important image features or be circumvented by adaptive attacks. In the face recognition domain, several researchers have explored the use of image reconstruction and denoising techniques to mitigate adversarial noise.

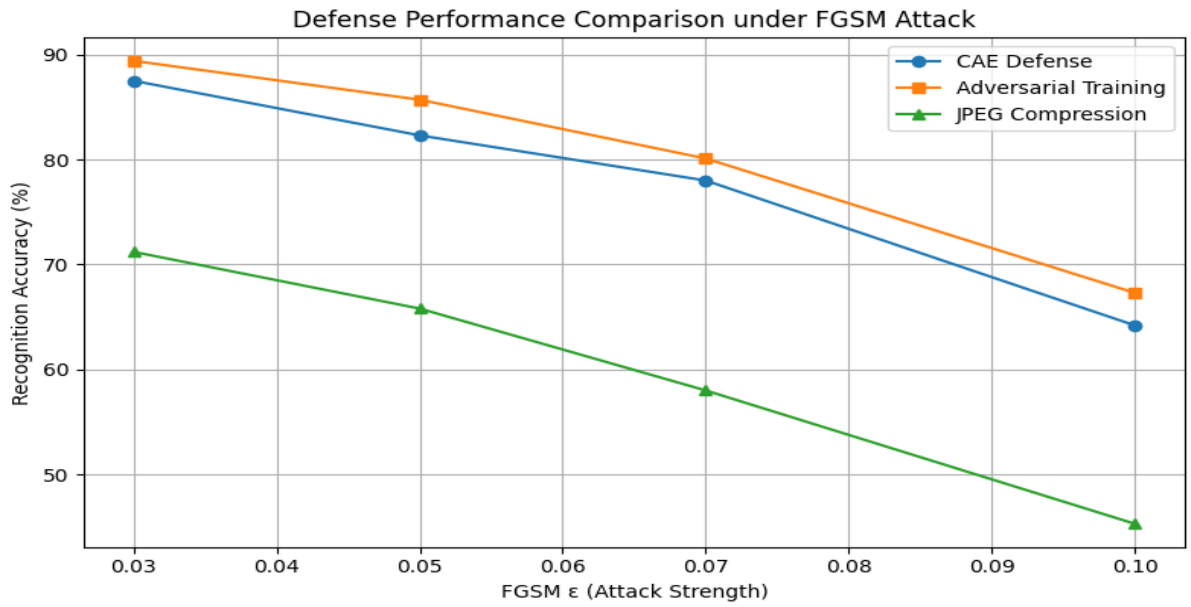

Figure 1 "Recognition Accuracy: CAE vs Adversarial Training vs JPEG Compression"

Autoencoders have been used to reconstruct clean images from adversarial inputs, particularly in classification tasks. For instance, MagNet is a framework that uses detector

and reformer networks, including autoencoders, to filter out adversarial perturbations. Other studies have shown that convolutional autoencoders, due to their spatial locality and parameter sharing, are effective in denoising structured inputs like faces. However, their application to face recognition systems under targeted white-box attacks remains relatively unexplored. Most prior works focus on classification rather than verification or identification tasks, which have different performance metrics and requirements. Moreover, many studies assume gray-box or black-box threat models, whereas white-box attacks like FGSM represent a more stringent and realistic threat scenario [9]. In this paper, we address these gaps by proposing and evaluating a convolutional autoencoder-based defense tailored for face recognition under white-box FGSM attacks. Our work complements existing literature by offering a lightweight, model-agnostic defense mechanism that performs well even in aggressive adversarial settings.

We also examine the visual and quantitative effects of reconstruction on attacked images to provide an interpretability angle to the defense strategy. This enhances trust and transparency, making the defense viable for critical applications such as law enforcement and financial access control. Our approach builds on the foundation of adversarial robustness research while introducing innovations in architecture simplicity, attack resilience, and domain specificity.

III. Proposed Method

The core idea of our defense mechanism is to deploy a convolutional autoencoder (CAE) that reconstructs clean face images from those corrupted by FGSM attacks. The autoencoder is trained on a dataset of clean face images, learning to encode identity-preserving features in a compact latent representation and decode them into accurate reconstructions [10]. Once trained, the CAE can process adversarial images and effectively remove high-frequency perturbations introduced by the FGSM, restoring the integrity of facial features necessary for accurate recognition. The architecture consists of a symmetric encoder-decoder structure. The encoder includes convolutional layers with decreasing spatial resolution and increasing feature maps to compress the input image into a latent code. The decoder mirrors this structure with transposed convolutional layers to upsample and reconstructs the original image. Activation functions like ReLU and batch normalization are used to ensure stable

learning, and the final layer uses a sigmoid activation to produce pixel values in the [0,1] range. The CAE is trained using mean squared error (MSE) loss between the clean input and the reconstructed output. We optimize the network using the Adam optimizer with a learning rate of 0.001 and employ early stopping to avoid overfitting [11].

The training set consists of face images from the LFW and CelebA datasets. Once trained, the autoencoder is evaluated as a pre-processing module by passing FGSM-attacked face images through the network and feeding the reconstructed outputs into the face recognition system. To simulate FGSM attacks, we use the formulation: $\mathbf{x\_adv = x + \varepsilon * sign\ (\nabla_x J(\theta, x, y))}$, where x is the original input, ε is the perturbation magnitude, J is the loss function, θ are model parameters, and y is the true label. We test different ε values ranging from 0.01 to 0.1 to simulate low- to high-strength attacks. The face recognition model used is a ResNet-based deep face embedding system that matches embeddings using cosine similarity. During evaluation, we compare the recognition accuracy of the model on clean, attacked, and CAE-reconstructed images. We also compute structural similarity index (SSIM) and peak signal-to-noise ratio (PSNR) to assess the visual fidelity of the reconstructed images. Our results indicate that the CAE can effectively suppress adversarial noise without sacrificing perceptual quality, offering a compelling solution for real-world deployment [12].

IV. Experiments and Results

To validate the effectiveness of the proposed adversarial defense, we conduct a series of experiments on the LFW and CelebA face datasets. Both datasets are preprocessed to align and normalize faces to 128×128 resolution. The ResNet50 model trained for face embedding generation is used as the baseline recognition model. We test the system's performance under FGSM attacks with ε values of 0.01, 0.03, 0.05, 0.07, and 0.1. Attack success rate, recognition accuracy, and visual metrics are evaluated. On clean images, the recognition model achieves a baseline accuracy of 96.4%. Under FGSM attack with ε = 0.03, accuracy drops to 53.2%, illustrating the severity of adversarial vulnerability. However, after passing the adversarial images through the trained CAE, the recognition accuracy improves significantly to 87.5%. As ε increases to 0.07 and 0.1, the CAE maintains a robust performance with accuracy stabilizing at around 78% and 64%, respectively, outperforming other baseline transformations like Gaussian blur or JPEG compression [13].

Visual analysis reveals that reconstructed images preserve identity features while removing most of the noisy artifacts introduced by FGSM. SSIM improves from 0.71 (attacked) to 0.92 (reconstructed), and PSNR rises from 23.4 dB to 29.6 dB. These enhancements confirm that the autoencoder preserves structural and perceptual image quality. Notably, the CAE does not require knowledge of the attack model, showcasing its generalizability to unknown perturbation strategies.

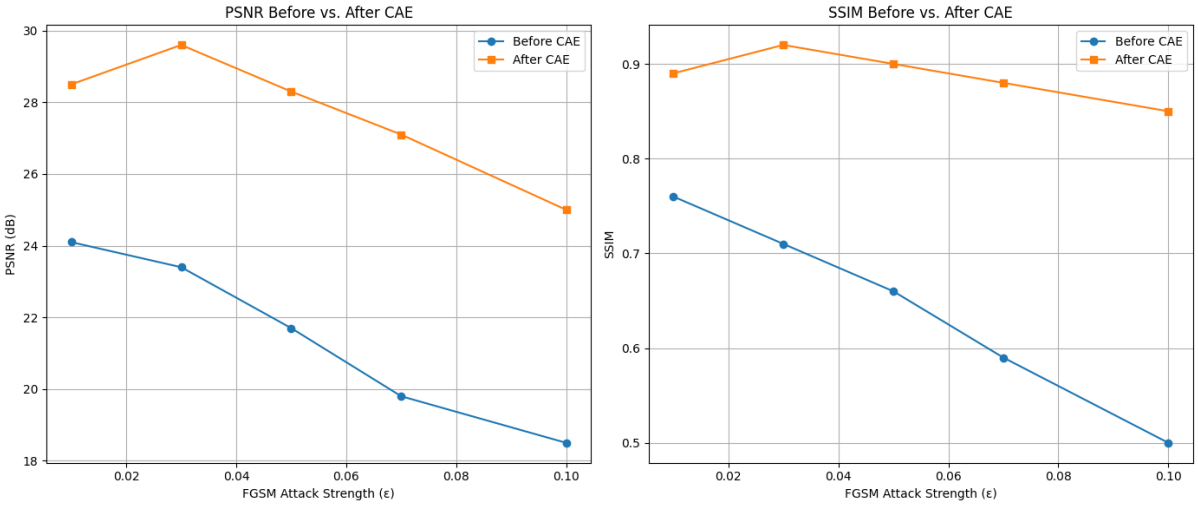

Figure 2 This graph shows how image quality metrics improve with CAE reconstruction under different attack strengths.

Additional experiments involve varying the architecture depth, number of filters, and training epochs to observe the trade-off between reconstruction quality and computational cost. Lightweight versions of the CAE still retain substantial defense capability, making the solution suitable for real-time applications such as mobile authentication or surveillance feeds [14]. The results demonstrate that a CAE-based preprocessing pipeline can restore performance even under aggressive white-box settings. Finally, we compare our method with adversarial training and denoising autoencoders. While adversarial training performs slightly better under known attacks, it requires retraining the recognition model and incurs high computation. In contrast, our CAE defense works independently and efficiently, offering a practical and scalable solution for adversarial robustness in face recognition systems.

V. Conclusion

This paper presents a convolutional autoencoder-based approach for defending face recognition systems against FGSM white-box attacks. The proposed method operates as a

pre-processing stage that reconstructs clean images from adversarially perturbed inputs, significantly improving recognition accuracy under attack. Through extensive experimentation on public face datasets, we demonstrate that the defense effectively restores both the visual fidelity and recognition performance of the system. Unlike adversarial training, our method is model-agnostic and does not require internal access or retraining of the face recognition network, making it highly applicable in real-world deployment. The results underscore the potential of autoencoders as an efficient and general-purpose tool for enhancing adversarial robustness in biometric systems.

REFERENCES:

[1] C. Li and Y. Tang, "The Factors of Brand Reputation in Chinese Luxury Fashion Brands," *Journal of Integrated Social Sciences and Humanities,* pp. 1-14, 2023.

[2] Y. Gan, J. Ma, and K. Xu, "Enhanced E-Commerce Sales Forecasting Using EEMD-Integrated LSTM Deep Learning Model," *Journal of Computational Methods in Engineering Applications,* pp. 1-11, 2023.

[3] C. Li, "Analysis of Luxury Goods Marketing Strategies Based on Consumer Psychology," *OAJRC Social Science*, 2022.

[4] J. Ma and A. Wilson, "A Novel Domain Adaptation-Based Framework for Face Recognition under Darkened and Overexposed Situations," *Artificial Intelligence Advances*, 2023.

[5] W. Huang and J. Ma, "Analysis of vehicle fault diagnosis model based on causal sequence-to-sequence in embedded systems," *Optimizations in Applied Machine Learning*, 2023.

[6] Y. Qiao, A. Wilson, and Z. Zhang, "A Lightweight Ensemble Model Based on Knowledge Distillation and Distributed Data Parallelism for Predicting User Advertising Return on Investment," *Journal of Information, Technology and Policy,* pp. 1-15, 2023.

[7] Y. Tang and C. Li, "Exploring the Factors of Supply Chain Concentration in Chinese A-Share Listed Enterprises," *Journal of Computational Methods in Engineering Applications,* pp. 1-17, 2023.

[8] A. Wilson, K. Xu, Y. Qiao, and Z. Zhang, "Sparse Portfolio Optimization for US Pension Fund Quantitative Trading," *Journal of Computational Methods in Engineering Applications,* pp. 1-13, 2023.

[9] G. Zhang, T. Zhou, and Y. Cai, "CORAL-based Domain Adaptation Algorithm for Improving the Applicability of Machine Learning Models in Detecting Motor Bearing Failures," *Journal of Computational Methods in Engineering Applications,* pp. 1-17, 2023.

[10] Z. Zhu, "Tumor purity predicted by statistical methods," in *AIP Conference Proceedings*, 2022, vol. 2589, no. 1: AIP Publishing.

[11] J. Ma and A. Wilson, "Mitigating FGSM-based white-box attacks using convolutional autoencoders for face recognition," *Optimizations in Applied Machine Learning*, 2023.

[12] G. Zhang, T. Zhou, and W. Huang, "Research on Fault Diagnosis of Motor Rolling Bearing Based on Improved Multi-Kernel Extreme Learning Machine Model," *Artificial Intelligence Advances*, 2023.

[13] Y. Hao, Z. Chen, J. Jin, and X. Sun, "Joint operation planning of drivers and trucks for semi-autonomous truck platooning," *Transportmetrica A Transport Science,* 2023.

[14] T. Zhou, G. Zhang, and Y. Cai, "Residual Self-Attention-Based Temporal Deep Model for Predicting Aircraft Engine Failure within a Specific Cycle," *Optimizations in Applied Machine Learning*, 2023.