Detection and Mitigation of Multimodal Adversarial Attacks

Dejaun Gayle (gayledej@kean.edu), Dr. Yulia Kumar (Advisor)

Department of Computer Science and Technology, Kean University, Union, NJ, USA

Abstract

In the rapidly evolving field of AI, ensuring the robustness of Large Language Models (LLMs) against adversarial attacks is mandatory. This study investigates the impact of combining adversarial images with regular images on targeted LLM's outputs and proposes strategies to combat these vulnerabilities. Utilizing techniques such as FGSM and PGD, we generated adversarial images and assessed their effects on models like ChatGPT and DALL-E. Our findings reveal significant vulnerabilities, highlighting the need for robust defense mechanisms. We propose adversarial training, defensive distillation, input preprocessing, ensemble methods, and regularization as mitigation strategies.

Introduction

Artificial Intelligence (AI) systems, particularly LLMs such as ChatGPT and DALL-E, are susceptible to Adversarial Attacks (AA). These attacks manipulate input data to deceive models, causing misclassifications or incorrect outputs. This research focuses on detecting adversarial images and developing methods to revert them, thereby enhancing the security of LLMs.

Background

AA involve creating inputs designed to fool AI models. Techniques such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are commonly used to generate adversarial examples. These alterations can significantly affect the performance of AI models, leading to incorrect or harmful outputs. Detecting and reverting them is crucial for maintaining the integrity and reliability of AI systems.

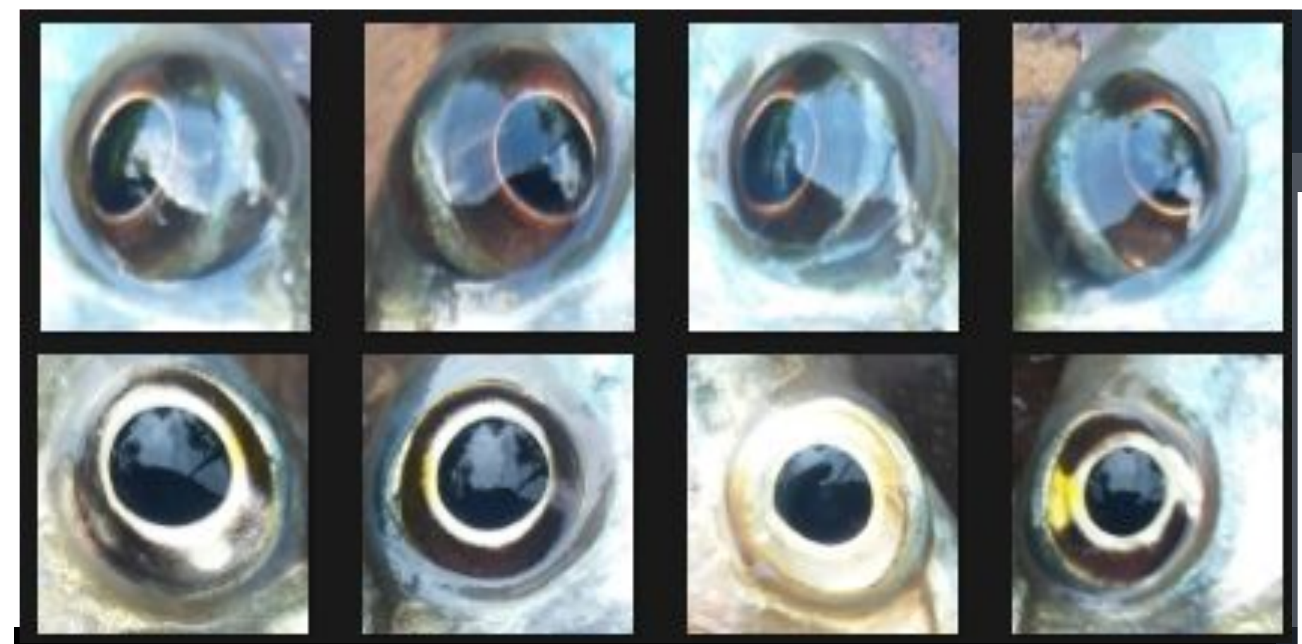

Figure 1. Fisheye Dataset:
(Fresh (Top) vs. non-fresh (Bottom)

Reference(s)

Materials and Methods

Data Collection: Datasets used include CIFAR-10, ImageNet, and custom datasets (Fisheye Dataset) containing both regular and adversarial images.

Adversarial Image Generation: Adversarial examples were generated using FGSM and PGD techniques. These methods involve adding small perturbations to input images to create adversarial examples. Figure 2 displays the original image and adversarial examples generated with varying epsilon values using the Fast Gradient Sign Method (FGSM). As epsilon increases, perturbations become more pronounced, and the model's confidence in its prediction drops or changes entirely. Figure 3 shows the relationship between the perturbation magnitude (epsilon) and the model's confidence. As epsilon increases, the model's confidence decreases, demonstrating the effectiveness of adversarial attacks.

Detection Techniques:

- Feature Squeezing: Reduces color bit depth and uses spatial smoothing to detect adversarial changes.
- Local Intrinsic Dimensionality (LID): Estimates local dimensionality to identify anomalies.
- Neural Network Ensembles: Uses multiple models to detect adversarial examples by comparing outputs.

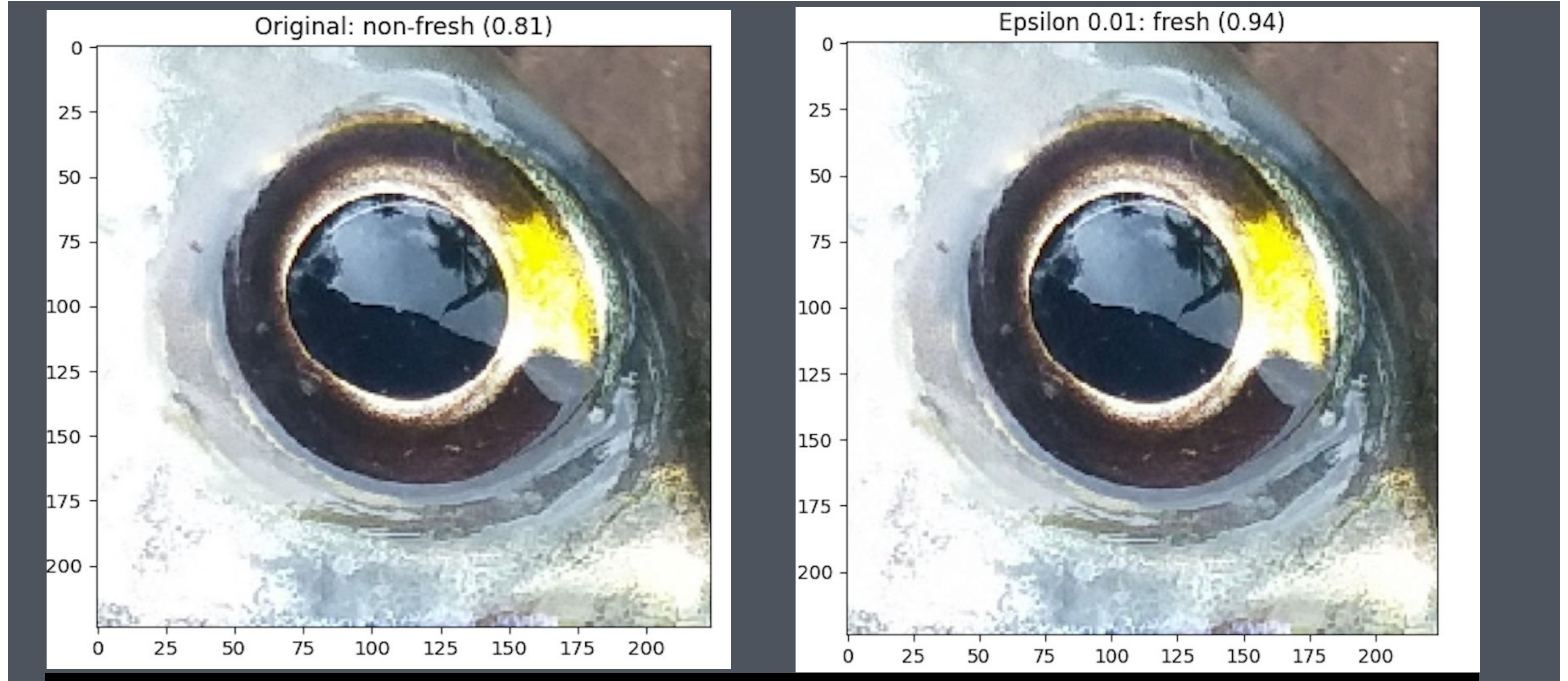

Figure 2. Original image input (left) vs altered FreedomGen app Output (right)

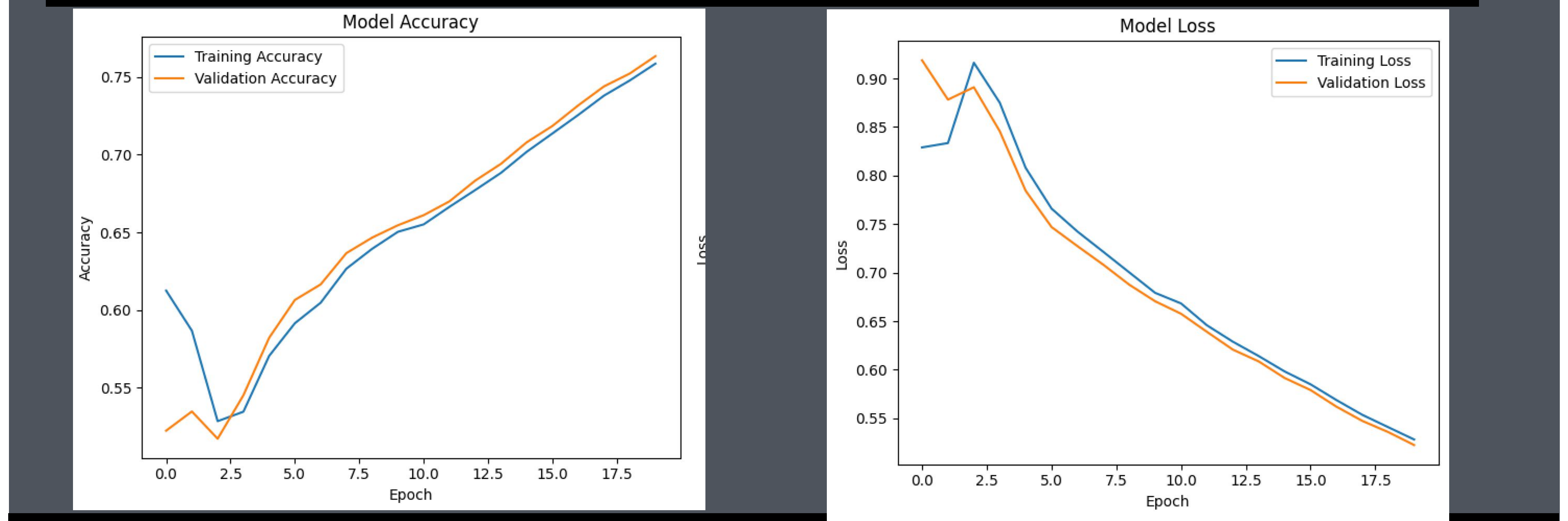

Figure 3: Impact of Epsilon on Model Confidence in Adversarial Attacks

Summary / Conclusion

The researchers developed their own FreedomGen app devoted to detection and mitigation of multimodal adversarial attacks on LLMs. The proposed FGSM and PGD techniques significantly enhance the security and robustness of AI models. Future research should explore additional detection methods, refine reversion techniques, and test on a broader range of models and datasets.

This material is based upon work supported by the National Science Foundation under Grants numbers CNS-2137791, and HRD-1834620. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.