Backdoor Attack Defense Report

Submitted By: *Priyanshi Singh (ps4609)*

GITHUB REPO: https://github.com/singh-priyanshi/backdoor-detector_for_BadNets

Overview

This project focuses on the implementation of a defense mechanism against backdoor attacks on machine learning models. The defense strategy utilizes a pruning-based approach to fortify the model against potential manipulations by backdoor patterns. The analysis involves the assessment of model vulnerability, identification of susceptible layers, selective pruning, and continuous evaluation of defense effectiveness.

Introduction

Backdoor attacks pose a significant threat to the integrity of machine learning models, allowing adversaries to manipulate model predictions by injecting subtle patterns. This project aims to develop a robust defense mechanism to detect and mitigate backdoor attacks, ensuring the model's reliability and security.

Methodology

1. Dataset

The project utilizes clean and backdoored datasets in HDF5 format. The datasets are loaded and preprocessed, ensuring compatibility with the model's input requirements.

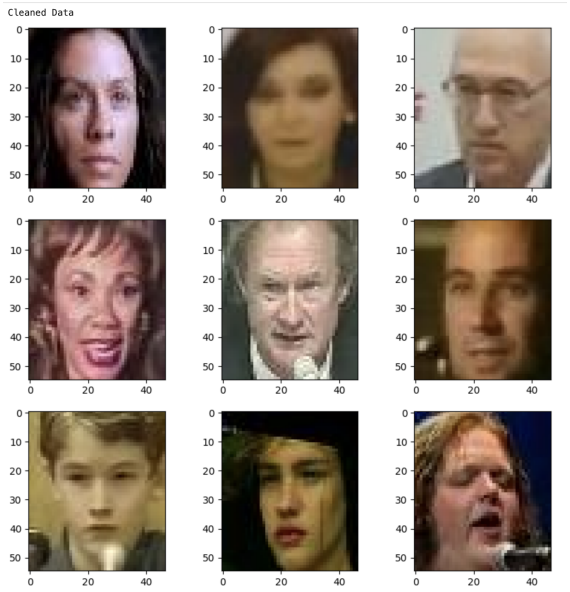
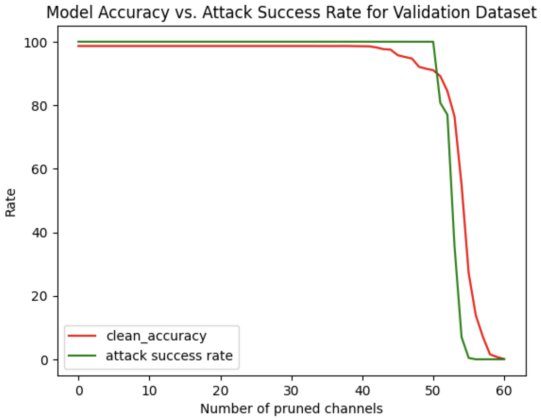


Fig: Cleaned Data

Fig: Poisoned Data

2. Model

A pre-trained model (`bd_net.h5`) serves as the target for backdoor attacks. The architecture of the model involves convolutional layers, max-pooling, and fully connected layers.



3. Pruning-Based Defense

The defense strategy consists of several key steps:

a. Backdoor Data Assessment: Evaluate the model's initial susceptibility to backdoor attacks by measuring the success rate on the backdoor dataset.

b. Identification of Vulnerable Layers: Identify layers prone to backdoor attacks, typically those close to the input where the backdoor pattern has a substantial impact.

c. Selective Pruning of Activation Channels: Prune specific activation channels in identified layers based on predefined criteria, such as mean activation values.

d. Reassessment of the Model: Reevaluate the pruned model on both clean and backdoor datasets after each round of pruning. Monitor changes in accuracy and attack success rate.

e. Iterative Pruning Process: Iterate the pruning process, gradually eliminating more channels or layers exhibiting characteristics of backdoor activation.

f. Saving Based on Thresholds: Save the pruned model if the accuracy on clean data falls below a specified threshold, indicating potential removal of the backdoor.

g. Continuous Evaluation: Perpetually monitor the effectiveness of the model's defense against backdoor attacks. Adjust pruning strategies, thresholds, or other parameters based on ongoing evaluations.

4. Model Performance Evaluation

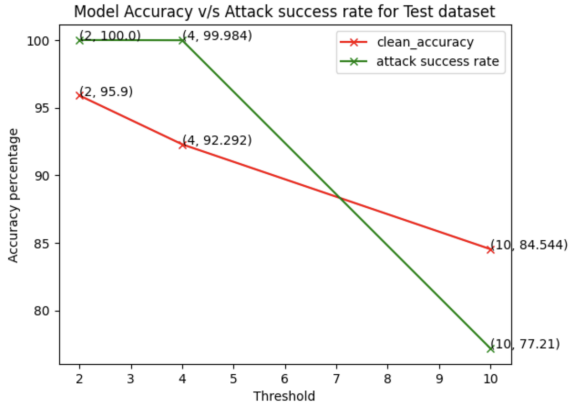
Explore the performance of the model on the validation dataset and visualize the clean and poisoned data. Calculate accuracy and attack success rate before and after the pruning defense to assess its impact.

5. GoodNet Model

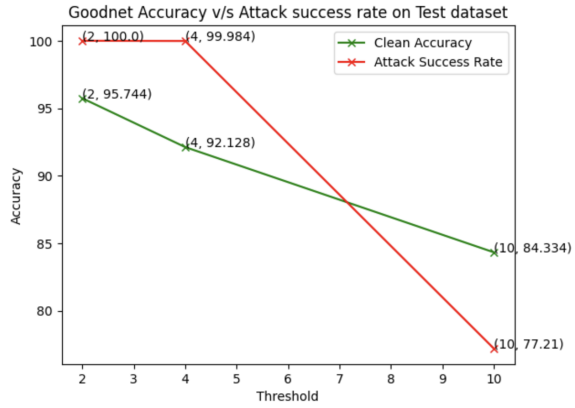
Introduce the GoodNet model, a combination of the original Backdoor Model (Bad Net) and the Repaired Backdoor Model. GoodNet aims to detect anomalies by comparing predictions from both models.

Results and Analysis

The project results reveal the effectiveness of the pruning-based defense strategy. The model's accuracy on clean data remains high, while the attack success rate on backdoor data decreases significantly after pruning. Visualizations provide insights into the model's behavior and the impact of the defense strategy.



The introduction of the GoodNet model further enhances the detection capability against backdoor attacks. By leveraging the agreement or disagreement between the original and repaired models, GoodNet identifies anomalies and provides an additional layer of security.



Analysis

model	test_accuracy	attack_rate
defence_model_2%	95.900	100.000
defence_model_4%	92.292	99.984
defence_model_10%	84.544	77.210

model	test_accuracy	attack_rate
repaired_model_2%	95.744	100.000
repaired_model_4%	92.128	99.984
repaired_model_10%	84.334	77.210

Conclusion

In conclusion, the developed defense mechanism showcases promising results in mitigating the impact of backdoor attacks on machine learning models. The iterative pruning process, coupled with continuous evaluation, contributes to a dynamic defense strategy. The GoodNet model introduces an innovative approach to anomaly detection, further enhancing the model's resilience against adversarial attacks.