Machine Learning for Cybersecurity
Fall 2023
Lab 4: Backdoor Attacks

Title: Design and Evaluation of a Pruning-Based Backdoor Detector for BadNets

Author: Noel Nebu Panicker

Abstract
This project focuses on developing a backdoor detector for BadNets, particularly those trained on the YouTube Face dataset, using a pruning-based defense mechanism. The detector assesses a neural network's susceptibility to backdoor attacks by pruning its last pooling layer and observing the impact on classification accuracy and attack success rates.

Introduction

- Objective: The goal is to implement a backdoor detector that utilizes a pruning defense strategy against BadNets, a type of neural network compromised by backdoor attacks.
- Background: BadNets are neural networks that contain hidden backdoors, allowing them to perform normally on regular inputs but act maliciously on specific, attacker-chosen inputs. Identifying and mitigating these backdoors is crucial for ensuring network security and integrity.

Methodology

- Dataset: The YouTube Face dataset serves as the primary data source for training and validation.
- Pruning Defense Approach: The pruning strategy involves systematically removing channels from the last pooling layer of the BadNet based on average activation values, and monitoring the drop in validation accuracy.
- Implementation Details: The neural network undergoes pruning until the validation accuracy drops by pre-defined thresholds (2%, 4%, and 10%). This process generates three variants of the original network, each corresponding to a different pruning level.

Graphical Representation of pruning process:

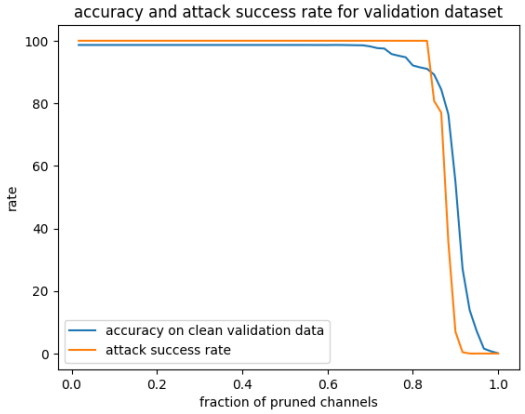


Table of Results:

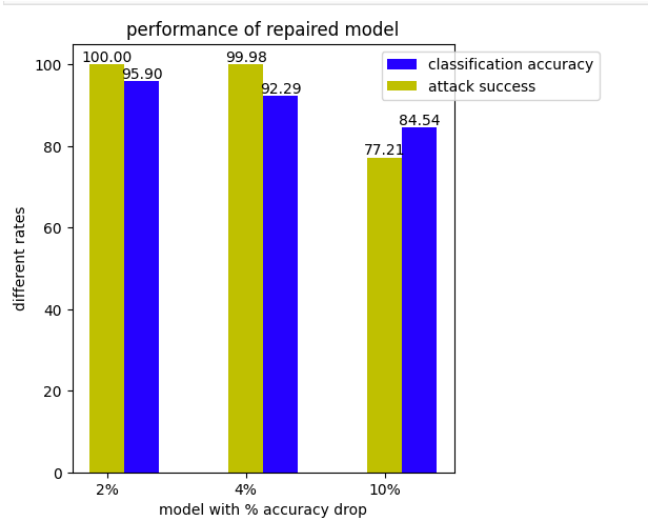
Results for pruned models on clean test data:

Pruning Percentage	Accuracy on Clean Test Data	Attack Success Rate
2%	95.90	100.0
4%	92.29	99.99
10%	84.54	77.21

Results for GoodNet models on clean test data:

Model Config	Accuracy on Clean Test Data	Attack Success Rate
2%	95.74	100.0
4%	92.12	99.98
10%	84.33	77.01

Graphs:



Discussion

The pruning defense's effectiveness is evaluated by its impact on maintaining classification accuracy for clean data while reducing the attack success rate. The optimal pruning level is determined based on the best balance achieved between these two metrics.

Conclusion

This project demonstrates the potential of pruning-based defenses in detecting and mitigating backdoor attacks in neural networks. The findings suggest that careful calibration of pruning levels is critical in preserving the network's functionality while enhancing its security.