ECE-GY 9163
ML for Cyber Security
Fall 2023

Designing and Evaluating a Pruning-Based Backdoor Detector for Neural Networks

Nagharjun M (nm4074)

12-04-2023

1 Introduction

This report contains explanations for the findings in Lab 4 - Designing and Evaluating a Pruning-Based Backdoor Detector for Neural Networks. Pruning defence is applied to a compromised BadNet model trained on the YouTube Face dataset, aiming to disable the backdoor while preserving accuracy for clean inputs. Further findings, based on various pruning levels, provide insights into balancing network integrity with robustness against backdoor attacks.

2 Dataset

The YouTube Face dataset is used in this experiment which is further divided into clean and poisoned subsets. I use clean validation and test sets (valid.h5 and test.h5) to fine-tune and assess the model. For backdoor simulation, poisoned datasets with a sunglasses trigger (bd_valid.h5 and bd_test.h5) are used to test the defence.

3 Workflow

I implemented a pruning defense on a neural network, selectively removing channels based on activation levels until accuracy dropped by predefined thresholds (2%, 4%, 10%). Using TensorFlow and Keras, I evaluated the pruned models against both original and poisoned data, classifying inputs as clean or backdoored based on model agreement. This approach aimed to balance accuracy with effective backdoor detection.

4 Results

Model	Repaired Clean Accuracy	Attack Rate
2% Repaired	95.7443	100
4% Repaired	92.1278	99.9844
10% Repaired	84.3336	77.2097

Table 1: Accuracy and Attack Rate of Repaired Models

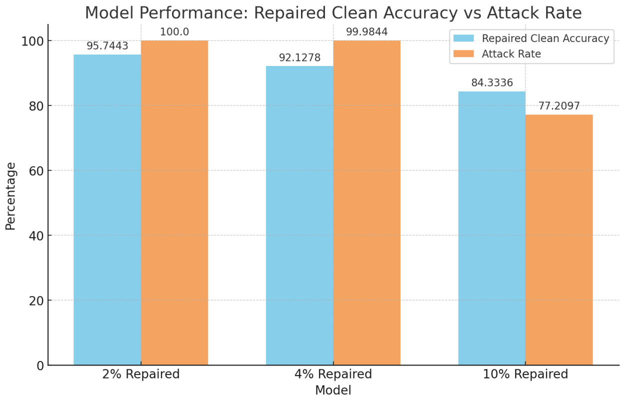

Figure 1: Comparison of Repaired Clean Accuracy and Attack Rate for each pruned model.

The results show that as the pruning threshold increases from 2% to 10%, the repaired model's accuracy on clean validation data decreases. This shows that there is a trade-off between accuracy and removing the backdoor. So, the thresholding should be chosen properly.