Analysis and Evaluation of Backdoor Detection in Cybersecurity Models - BadNet

Avinash Vijayarangan (av3134)

Abstract. This lab project delves into the innovative application of pruning strategies for enhancing the security of neural networks, particularly focusing on backdoor detection in cybersecurity models. Utilizing the YouTube Face dataset, the study investigates the efficacy of a Pruning-Based Backdoor Detector implemented on a compromised neural network model, known as BadNet. The research primarily aims to neutralize embedded backdoors without significantly impacting the accuracy of the model when processing clean data inputs. Through a series of experiments, various pruning intensities are meticulously examined to strike a critical balance between ensuring robust network security and maintaining optimal performance. The findings of this study provide valuable insights into the effectiveness of pruning as a defense mechanism against backdoor attacks in neural networks, highlighting the trade-offs between model accuracy and security. This work contributes to the growing field of cybersecurity, offering a novel perspective on safeguarding neural networks against sophisticated backdoor threats.

1 Introduction

This study discusses the outcomes of Lab 4, focusing on a Pruning-Based Backdoor Detector for Neural Networks. The research explores using a pruning defense on a compromised neural network model (BadNet), trained using the YouTube Face dataset. The objective is to neutralize the backdoor without compromising the accuracy of clean data inputs. Various pruning intensities were tested to find a balance between network security and performance.

2 Dataset

The YouTube Face dataset, comprising both clean and poisoned data, was utilized. The clean validation and test datasets (valid.h5 and test.h5) were used for model fine-tuning and evaluation. For backdoor attack simulations, poisoned datasets with a sunglasses trigger (bd valid.h5 and bd test.h5) were employed.

3 Workflow

The study involved implementing a pruning defense strategy on a neural network. This involved selectively removing network channels based on their activation levels until the accuracy fell by predefined thresholds (2%, 4%, 10%). Using TensorFlow and Keras, the performance of pruned models was assessed against both original and poisoned data. The method aimed to maintain accuracy while effectively detecting and mitigating backdoor threats.

4 Results

The study revealed that increasing the pruning threshold from 2% to 10% inversely affected the model's accuracy on clean validation data. This highlighted a compromise between maintaining accuracy and eliminating the backdoor. The results emphasized the importance of carefully selecting the pruning threshold.

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

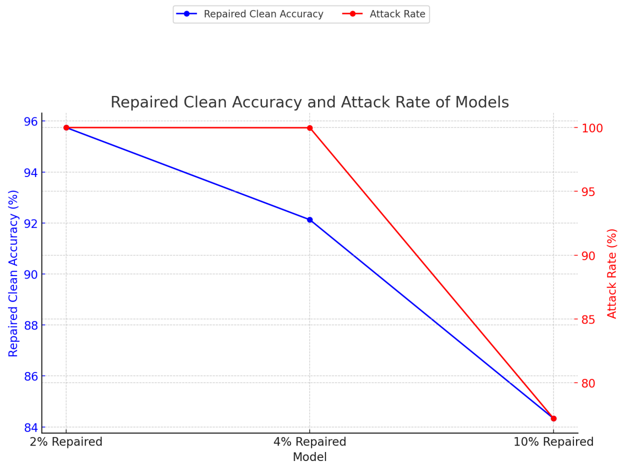

Fig 1 Comparison of Repaired Clean Accuracy and Attack Rate for each pruned model.