Backdoor Attacks - Report

Mannal Kamble - mk8475

December 4, 2023

Introduction

The objective of this project was to design a backdoor detector for BadNets trained on the YouTube Face dataset using the pruning defense. The goal was to create a repaired network, denoted as G, with $N + 1$ classes, where N is the number of original classes. The detector would output the correct class for clean inputs and class $N + 1$ for backdoored inputs.

Methodology

Pruning Defense

- The pruning defense involved removing one channel at a time from the last pooling layer of the BadNet B, which is just before the fully connected (FC) layers.
- Channels were pruned in decreasing order of average activation values over the entire validation set.
- Pruning continued until the validation accuracy dropped at least $X\%$ below the original accuracy, resulting in the new network B'.

Backdoor Detector G

- For each test input, the detector ran it through both the original BadNet B and the pruned BadNet B' (repaired network).
- If the classification outputs were the same (class i), the detector outputted class i. If they differed, the output was $N + 1$.

Evaluation

The defense was evaluated on the following scenarios:

BadNet

- A specific BadNet with a known "sunglasses backdoor" on the YouTube Face dataset.
- The detector's accuracy on clean test data and the attack success rate on backdoored test data were measured.

Repaired Networks for $X = \{2\%, 4\%, 10\%\}$

- The repaired networks (B') for different pruning levels (X) were evaluated using the evaluation script provided.
- Accuracy on clean test data and the attack success rate on backdoored test data were recorded.

Combined Network G'

The combined network G' was evaluated in a similar manner to assess its effectiveness in detecting backdoors.

Results

The table below summarizes the results:

Model	Repaired_2%	Repaired_4%	Repaired_10%	G_2%	G_4%	G_10%
Test Accuracy	95.90	92.29	84.54	95.74	92.13	84.33
Attack Rate	100.00	99.98	77.21	100.00	99.98	77.21

Conclusion

The results indicate that the pruning defense successfully repaired the BadNet for different levels of pruning. The backdoor detectors, G and G', demonstrated high accuracy on clean test data while effectively identifying backdoored inputs. The trade-off between accuracy and attack success rate was observed, with higher pruning levels leading to decreased attack success rates.

Conclusion

The results of the backdoor detection using the pruning defense reveal a notable trade-off between model accuracy on clean test data and vulnerability to backdoor attacks. While the pruning defense successfully repairs the network to some extent, evidenced by the decline in attack rate with increasing pruning levels, the backdoor remains highly effective.

Code and Repository

The code for this project is available on the GitHub repository: Lab 4 - Backdoor Attacks. The repository includes all necessary files, and the README provides instructions on how to run the code.