ML for CyberSecurity Lab4 Report - hn2237

Introduction

The objective of this lab is to design a backdoor detector for BadNets trained on the YouTube Face dataset using the pruning defense strategy discussed in class. The goal is to create a reliable defense mechanism that accurately identifies backdoored inputs while maintaining high accuracy on clean data.

Methodology

1. Pruning Defense

The first step involves applying the pruning defense to the given BadNet B. The pruning process targets the last pooling layer of BadNet B, removing one channel at a time in decreasing order of average activation values over the entire validation set. The pruning continues until the validation accuracy drops by at least X% below the original accuracy.

2. GoodNet G

The GoodNet G is designed to compare the classification outputs of both the original BadNet B and the pruned BadNet B'. For each test input, if the classification outputs of B and B' are the same (i.e., class i), the detector outputs class i. If the outputs differ, the detector outputs class N+1, indicating a potential backdoored input.

3. Repaired Networks for X={2%,4%,10%}

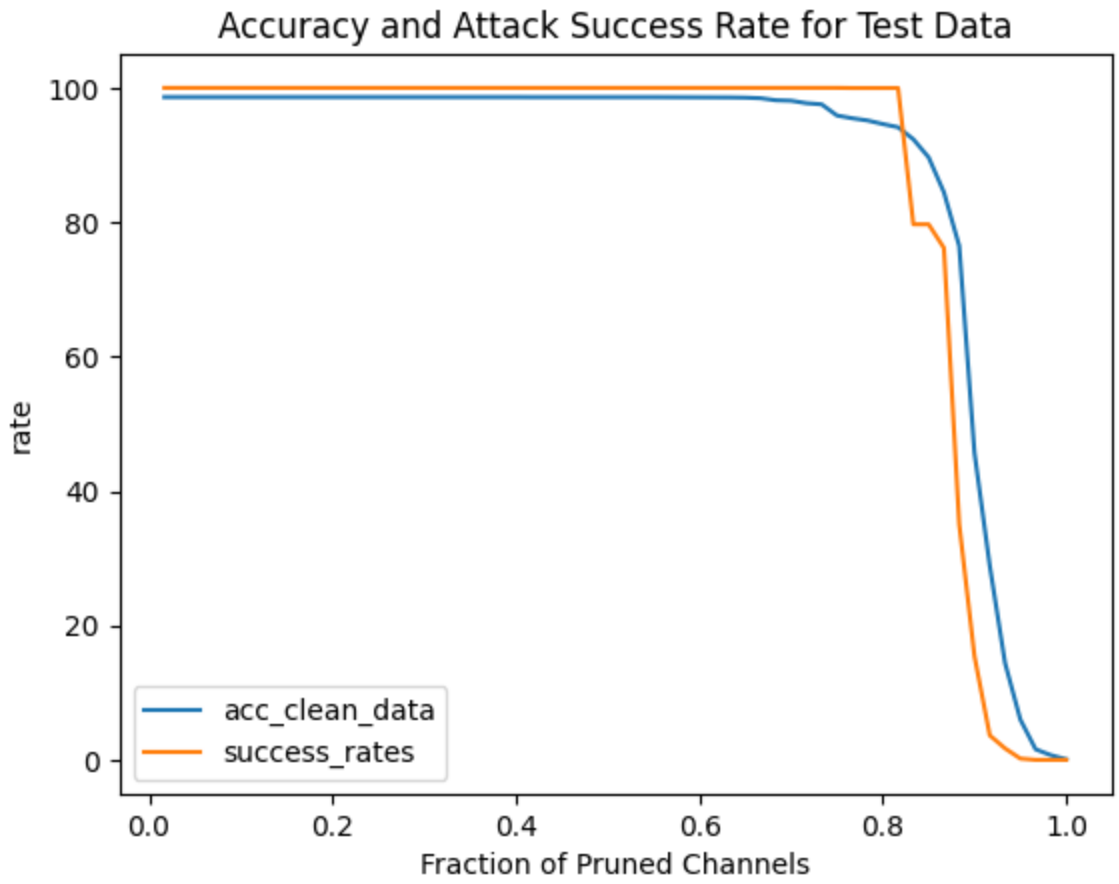
We focus on evaluating and testing the repaired networks for different pruning percentages (X={2%,4%,10%}).

Results

The following table summarizes the accuracy on clean test data and the attack success rate (on backdoored test data) as a function of the fraction of channels pruned (X):

Pruning Percentage (X)	Clean Data Accuracy	Attack Success Rate
2%	96.04053000779423 %	100.0 %
4%	94.81683554169913 %	99.97661730319564 %
10%	84.67653936087295 %	76.1730319563523 %

The following plot shows accuracy on clean test data and the attack success rate (on backdoored test data) as a function of the fraction of channels pruned (X):



We can observe that the defense is not too successful as the accuracy is sacrificed, when the pruning percentage (X) increases.

Code Repository

The code for this project can be found in the GitHub repository:
https://github.com/Soester10/NYU_ECE9163_MLSec_Lab4

Running the Code

Please refer to the README.md file in the repository for instructions on how to run the code and reproduce the results.

Conclusion

In conclusion, the backdoor detector based on the pruning defense strategy shows decent results in identifying backdoored inputs, however it doesn't maintain a high accuracy on clean data. Further improvements and optimizations can be explored for enhanced defense mechanisms in real-world scenarios, while maintaining a high accuracy.