Report

On

Backdoor Detection with Pruning Defense

Nishita Sinha

ns5418

INTRODUCTION

The objective of this project was to design a backdoor detector for BadNets trained on the YouTube Face dataset using the pruning defense strategy. The approach involved iteratively pruning channels from the last pooling layer of a backdoored neural network (BadNet B) based on decreasing average activation values, resulting in a repaired network (B'). The GoodNet G was then designed to compare classifications between B and B' for backdoor detection.

IMPLEMENTATION

The primary objective of this project was to enhance a machine learning model by undertaking a sequence of actions, such as layer pruning, saving models based on accuracy, conducting vulnerability assessments, and developing an optimized composite model.

Layer Pruning and Model Saving:

In the process of layer pruning and model preservation, our methodology included the pruning of the conv_3 layer based on the average activation derived from the last pooling operation across the validation dataset. We adopted a strategy to save models at defined accuracy drop thresholds of 2%, 4%, and 10%. These models were designated as model_X=2.h5, model_X=4.h5, and model_X=10.h5, respectively, indicating the corresponding percentage of accuracy drop.

Vulnerability Assessment:

The success rate of attacks under the condition that the model's accuracy decreased by a minimum of 30%. This metric was noted to be 6.954187234779596%, which proves the existence of a vulnerability threshold.

Model Integration:

To improve the model's effectiveness, a strategy was implemented to merge two models, namely "BadNet" and an enhanced model after undergoing repairs. This created an advanced integrated model which can be called “GoodNet”.

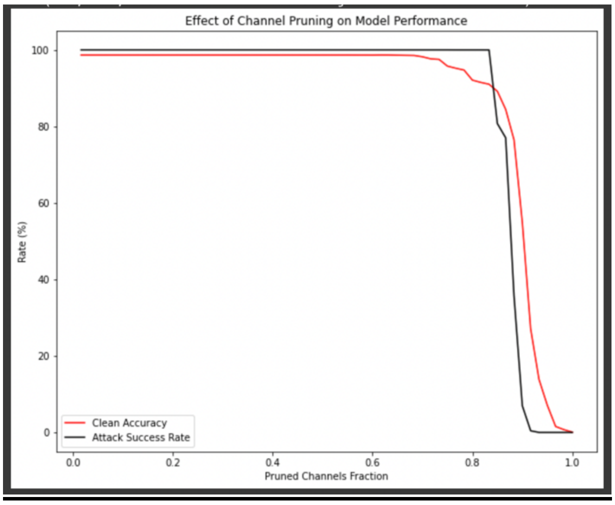
DATASET

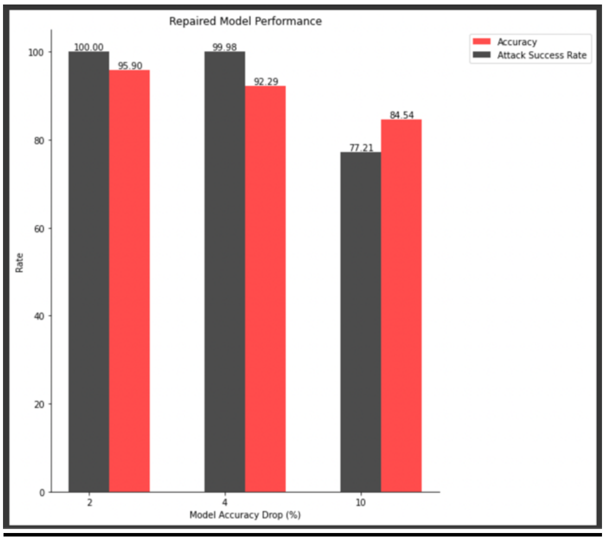
https://github.com/csaw-hackml/ CSAW-HackML-2020

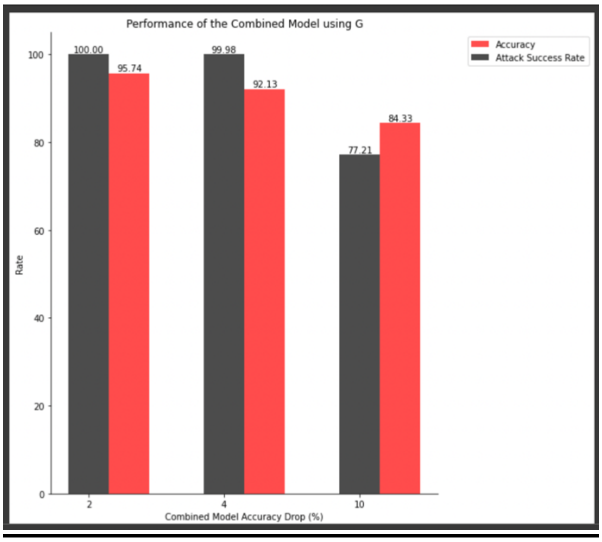
GITHUB LINK

https://github.com/nishitasinha24/ML_CyberSec_HW4/

RESULTS


Effect of Channel Pruning on Model Performance
Rate (%)
100
80
60
40
20
0
Clean Accuracy
Attack Success Rate
0.0
0.2
0.4
0.6
0.8
1.0
Pruned Channels Fraction


Repaired Model Performance
Accuracy
Attack Success Rate
Rate
100
80
60
40
20
0
100.00
95.90
99.98
92.29
77.21
84.54
2
4
10
Model Accuracy Drop (%)


Performance of the Combined Model using G
Accuracy
Attack Success Rate
Rate
100
80
60
40
20
0
100.00
95.74
99.98
92.13
77.21
84.33
2
4
10
Combined Model Accuracy Drop (%)

Repaired Model Performance

X(%)	Accuracy on Clean Test Data	Attack Success Rate
2	95.90	100.00
4	92.29	99.984412
10	84.54	77.209665

Performance of the Combined Model using G

X(%)	Accuracy on Clean Test Data	Attack Success Rate
2	95.90	100.00
4	92.29	99.984412
10	84.54	77.209665

CONCLUSION

The findings from the data underscore the inherent challenge in achieving a balance between maximizing model accuracy and bolstering resilience against potential attacks when implementing the pruning defense strategy. The exploration of alternative defense mechanisms or the thoughtful integration of multiple strategies may become imperative to forge a more effective and harmonized outcome.