REPORT

Design and Evaluation of a Pruning-based Backdoor Detector for BadNets

Github Link:https://github.com/kaladin1904/ML-For-CyberSecurity/tree/main
Name: Ishan Miglani
NetID: im2410

Abstract

This report outlines the development of a backdoor detector for BadNets, tailored to the YouTube Face dataset. Through a pruning defense approach, the detector repairs a backdoored neural network (BadNet) by selectively pruning channels within its last pooling layer. The resulting network, named GoodNet, categorizes inputs into their original classes or an additional class if detected as backdoored. This assignment addresses the pressing issue of backdoor attacks in neural networks, focusing on classifiers (BadNets) that operate normally on standard inputs but produce erroneous outputs when exposed to specific triggers.

Methodology

For this assignment, we employed the YouTube Face dataset, consisting of clean and backdoored validation and test images. Our foundation was a backdoored neural network, BadNet, which encompassed N=1283 classes. The defense strategy centered on pruning BadNet's last pooling layer, eliminating channels based on their average activation values across the validation set. We executed pruning in a descending order of these values. Subsequently, we designed GoodNet to leverage the outputs of both the original and pruned BadNets. When both networks concurred on a classification, GoodNet yielded that class. However, if their outputs diverged, GoodNet classified the input as class N+1, signifying a backdoored input.

Implementation

The data were loaded and prepared using TensorFlow and Keras. BadNet was then pruned iteratively, assessing changes in validation accuracy with each channel removal. Pruning stopped once the accuracy dropped by predefined thresholds (2%, 4%, 10%). Subsequently, GoodNets were constructed corresponding to each threshold.

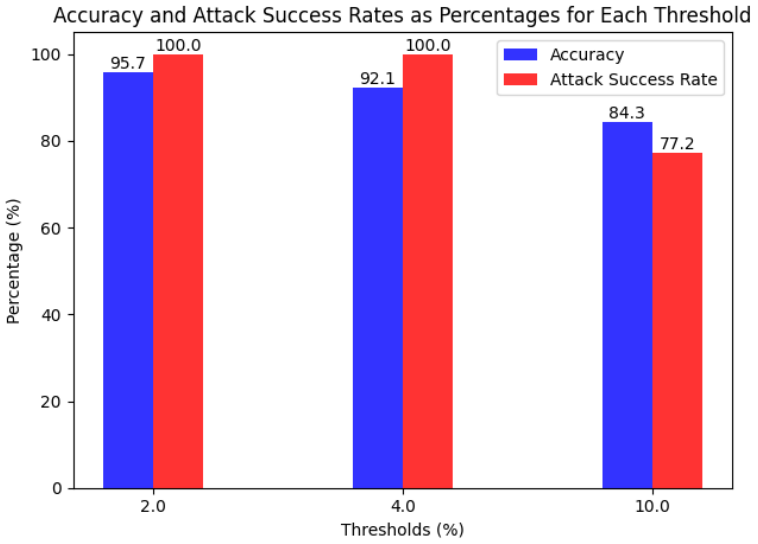
Results

Effects of Pruning

Pruning at different levels demonstrated a trade-off between maintaining accuracy on clean data and reducing the attack success rate. The accuracy on clean data decreased with increased pruning, while the ability to detect backdoored inputs improved.

Performance of GoodNet Models

Three GoodNet models were evaluated, corresponding to the pruning thresholds. The models showed high accuracies on clean test data (95.74%, 92.13%, and 84.33% respectively) and were effective in identifying backdoored test inputs, with detection rates of 100%, 99.98%, and 77.21% respectively. **For accuracy clean test data was used and for attack success rates backdoored test data was used.**



Threshold (%)	Accuracy (%)	Attack Success Rate (%)
2%	95.900234	100.000000
4%	92.291504	99.984412
10%	84.544037	77.209665

Performance of Pruned Models

It was observed that the 3 pruned models that the 3 goodnets were created from had the exact same results as the 3 Goodnet models shown above in terms of accuracy and attack success rates.

Interpretation

The results indicate that pruning can effectively reduce the success of backdoor attacks in neural networks. However, this comes at the cost of reduced accuracy on clean data, particularly at higher levels of pruning.

Conclusion

This assignment demonstrates the feasibility of using pruning as a defense against backdoor attacks in neural networks. The GoodNet models developed show promising results in both maintaining high accuracy on clean data and effectively detecting backdoored inputs.