A Backdoor Detector for BadNets trained on the YouTube Face dataset using the pruning defense

yl9393 Yiwen Liang

Methodology

In this project, first get the mean value of each channel's activation in the last pooling layer and sort them in increasing order and store the index. After that, prune the BadNet B convolution layer 3 by setting each channel's weight and bias zero in this layer one by one following the index order. And each time, after pruning the channel, measure the new validation accuracy of the new BadNet until the accuracy drops at least X below the original accuracy. Therefore, we got a new BadNet B_prime. After repairing these two networks, we can get a new N+1 classes network, and test these networks with clean dataset and sunglasses poisoned dataset, to get the accuracy result on clean test and the attack success rate.

Result

Below shows the result about the accuracy on clean test data and the attack success rate of different fraction of channels pruned (X)
For X = 2%

	accuracy on clean test data	attack success rate
B_prime	95.90	100.0
B	98.62	100.0
repaired_net	95.74	100.0

For X = 4%

	accuracy on clean test data	attack success rate
B_prime	92.29	99.98
B	98.62	100.0
repaired_net	92.12	99.98

For X = 10%

	accuracy on clean test data	attack success rate

B_prime	85.54	77.20
B	98.62	100.0
repaired_net	84.33	77.20

From above results, we found that by only pruning one layer in BadNet, we can defend against the sunglasses attack a little bit, but the attack success rate dropness is not obvious. Therefore, I assume that if pruning more layers in BadNet, there will be a better attack defense result, with a lower accuracy of recognition in clean test data.

GitHub repo

https://github.com/haborta/ML_CyberSecurity/tree/main/lab2