Back door attack and pruning defense

In this lab. I designed a backdoor detector for BadNets trained on the YouTube Face dataset using the pruning defense method.

Firstly, I have a BadNet model, and I have two datasets, one is the clean dataset with correct label, and the other contains bad dataset with targeted label. So, I can use the clean dataset to test the clean classification accuracy for the backdoor model, while using the bad dataset to get the attack success rate.

Originally, for the BadNet under testing in this lab, we got 98.62% for clean classification accuracy and 100% attack success rate. The attack success rate is higher than the clean classification accuracy, so the attack is really working.

Clean Classification accuracy for badnet	Attack success rate for badnet
98.62%	100%

Performance of original BadNet model and test data

Then I implemented the pruning defense in this BadNet. I pruned the last pooling layer of BadNet (the layer before the FC layers) by removing one channel at a time from that layer followed by increasing order of average activation values over the entire validation set.

Validation accuracy drop X%	Fraction of channels pruned	Clean classification accuracy	Attack success rate
2%	75%	95.74%	100%
4%	80%	92.13%	99.98%
10%	85%	84.33%	77.21%

Performance of the repaired_net on the clean test data

I repaired the network in three circumstances according to the percentage of validation accuracy drops 2%, 4% and 10%.

I found that when removing one channel which has a large average activation value over the entire validation set, the clean classification accuracy and attack success rate will drop more sharply. And when the validation accuracy dropped by 10%, the attack success rate also has a significant drop (almost 23%), but the clean classification accuracy reduce significantly as well. So I think for this backdoor attack, pruning is not very effective due to while the attack success rate of the attack decreased, the clean classification accuracy also decreased. In the future, maybe wo could try to use fine-pruning or other defense methods like STRIP to further defense backdoor attack in this model.