Neelanchal Gahalot
ng2436

Pruning Defense On Backdoored Networks

ECE-GY 9163 ML for Cyber Security Lab 2 Report

Problem Formulation and Description

We design a backdoor detector for BadNets trained on the YouTube Face dataset using the pruning defense.

As input, we take:

- a backdoored neural network classifier with N classes.
- a validation dataset of clean, labelled images.

As output, we have:

- Correct class if test input is clean, in [1,N].
- If backdoored, we output N+1 classes.

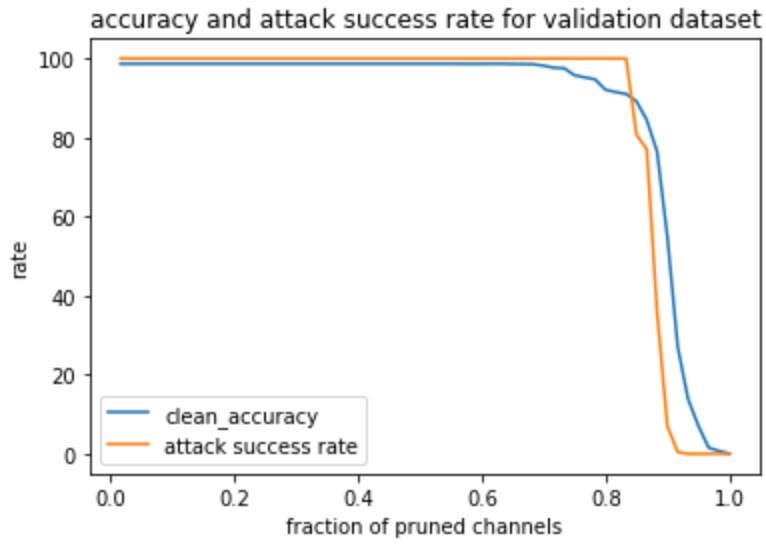
Methodology:

Pruning is essentially removing one channel at a time from a pooling layer.. We do that in decreasing order of average activation values over the entire validation set. We measure validation accuracy everytime we prune a channel and stop pruning once the accuracy drops least X% below the original.

Results:

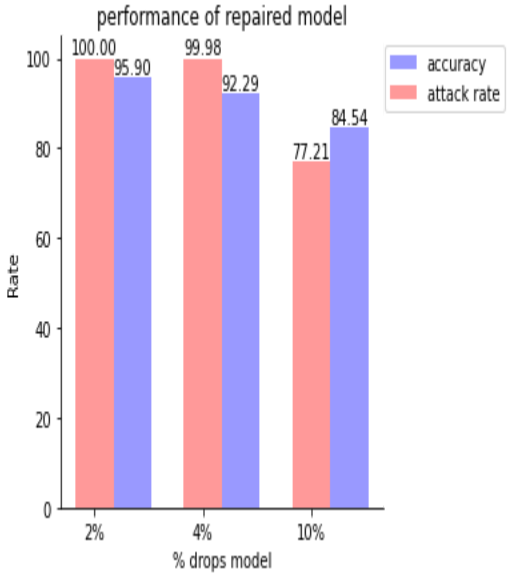
Three different Fixed BadNets with different levels of accuracy drop (2%, 4% and 10%) have been created (please refer to the notebook for more information). Changes in Clean Validation Accuracy and Attack Success Rate as a function of Fraction of Channels Pruned:

Neelanchal Gahalot
ng2436



Performance of our fixed model:

model	text_acc	attack_rate
repaired_2%	95.900234	100.000000
repaired_4%	92.291504	99.984412
repaired_10%	84.544037	77.209665



Conclusion:

Pruning offers a good defense against backdoor attacks but it is not very effective by itself. When combined with Tuning, it can offer even better results. Fine-pruning, a combination of pruning and fine-tuning, and show that it successfully weakens or even eliminates the backdoors, i.e., in some cases reducing the attack success rate to 0% with only a 0.4% drop in accuracy for clean (non-triggering) inputs[1].

References:

Neelanchal Gahalot
ng2436

[1] Liu, K., Dolan-Gavitt, B., & Garg, S. (2018, September). Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses* (pp. 273-294). Springer, Cham.