Lab3 report

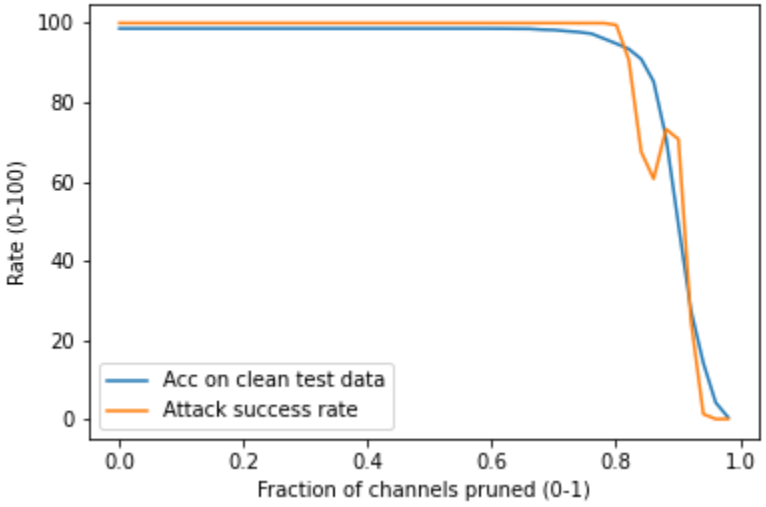
1. Evaluations of X = {2%, 4%, 10%, 30%}

Accuracy dropped in clean validation set (%)	Accuracy on clean test set (%)	Attach success rate on backdoored test set (%)
2	96.58612626656274	99.99220576773187
4	94.77786438035854	98.46453624318005
10	88.09041309431022	63.16445830085736
30	67.8332034294622	81.60561184723305

As more channels are pruned, the accuracy on the clean validation set and that on the clean test set drop at almost the same pace. The attack success rate on the backdoored test set first drops for a while, but at the last row it goes up again.

2. Accuracy on clean data and attack success rate curves

The figure is shown as below:



It drops first but then rises up again. After that, as the pruning fraction goes to near 1, which means all channels are pruned, the attack success rate and accuracy both drop quickly to 0.

3. Comments on whether the pruning defense works

Obviously the pruning defense doesn't work well. Pruning defense is based on the observation that the backdoor attacks often rely on the normally inactive channels. If these inactive channels are activated, it may indicate that the net is attacked by some backdoor behavior. However, this

assumption may fail if the attack doesn't come from the inactive channels, which will cause the pruning defense to fail.

In the MNIST dataset example shown in the lecture, the backdoor sometimes comes from the pixels at the corners, which is far away from the digits at the center. This manner follows the assumption pretty well, leading to the success of pruning defense. But in our case, there is no promise on this assumption in any sense, so the pruning defense may fail, just as we can observe from the plot.