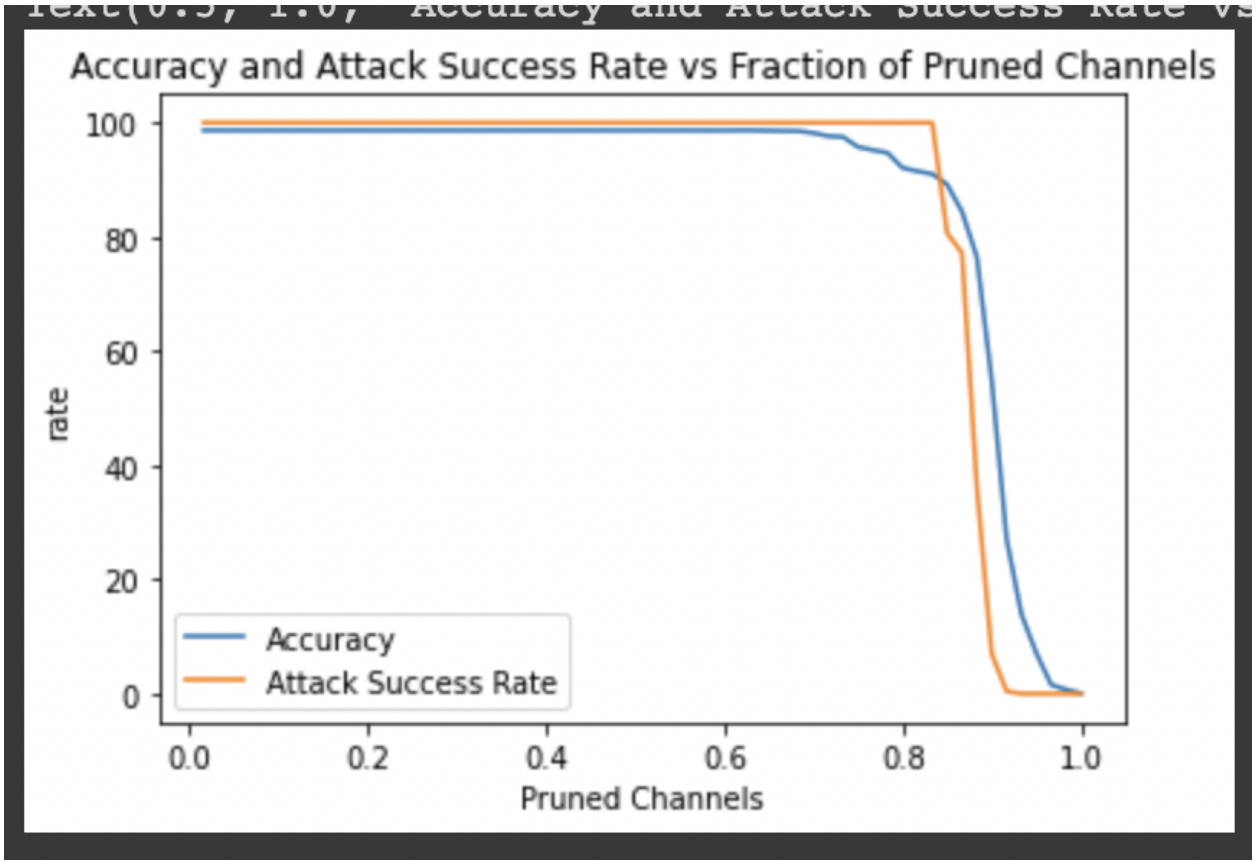
Pratap Narra pn2158

Designing a backdoor detector for BadNets trained on the Youtube Face Dataset using the pruning defense

https://github.com/pratapnarra/ML_CyberSec_HW2_Fall2022

So the network G is designed by pruning the last pool layer of the network, by removing one channel at a time from that network. Also, every time a channel is pruned valid accuracy of pruned badnet is measured, models are saved after the accuracy drops by 2%,4%, and 10%.

The below graph has the accuracy and attack success rate as a fraction of pruned channels, as we can clearly see as the fraction of pruned channels increase the attack success rate falls rapidly.



% of accuracy drop	Clean Accuracy	Attack Success Rate
2	95.90	100.0

4	92.29	99.98
10	84.54	77.209

From the table we can conclude that we are able to reduce the attack success rate significantly at the cost of a 10% drop in accuracy.