EL-GY-9163 ML Cyber Sec
Lab: Backoor attacks

sss9772@nyu.edu

- Methodology

1) Find the base accuracy and the base attack success rate of the model based on the clean dataset validation split.

2) Now find the activations at last pooling layer. At this layer there are 60 convolution channels and we need to prune them, in increasing order of the average activation over the validation dataset. This method is mentioned by Garg et. al. [Fine Pruning]

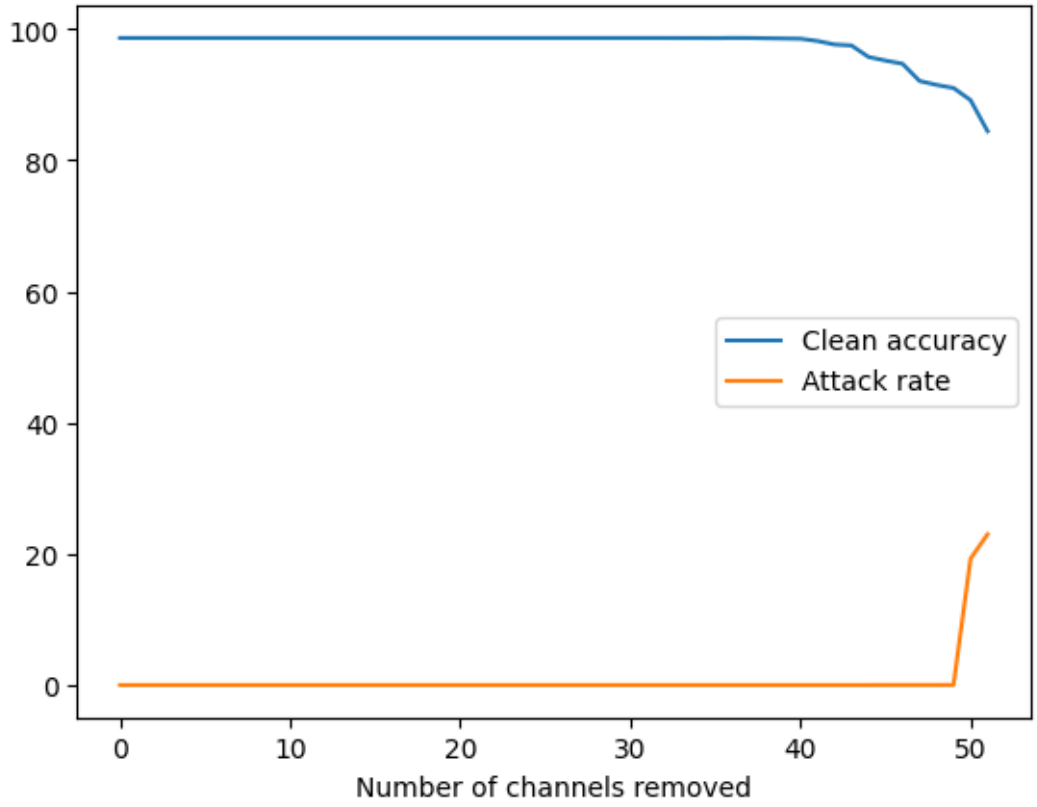
3) Sequentially prune a channels one by one, i.e. set the weights and bias of channel to zero. These channels are the least activated and based on these channels the attackers usually creates backdoors which leads the model to misclassify.

4) As we prune a channel each time, we measure the validation accuracy on the clean dataset and if it's below a certain threshold, we save the model as benchmark for the particular model accuracy.

5) Once we have different models created based on different model accuracies, we created a Goodnet out of the BadNet and BadNet_pruned. If the original badnet and the pruned badnet agree on a outcome, then the model prediction is given as it is. Otherwise, we say that the model just predicted upon a compromised bad dataset item. Hence the Goodnet has N+1 dimension on the last layer.

Results.

We observe that as we prune more channels, the validation accuracy drops. And after reaching a certain number of channels, the accuracy drops drastically. This is because once inactive channels are dropped, they have no effect but as soon as active channels are dropped, the model performs badly.



As far as the Good net accuracy is concerned, we see that the model accuracy drops just the accuracy of the badnet_pruned model drops, but we also see that the attack rate also increases, as this accuracy drops.

	Clean accuracy	Attack Rate
G_2	95.744349	0
G_4	92.127825	0.015588
G_10	84.333593	22.790335