ECE-GY 9163: Machine Learning for Cyber Security
Lab 4: Backdoor Detectors

Chandana Thimmalapura Jagadeeshaiah
ct3002@nyu.edu

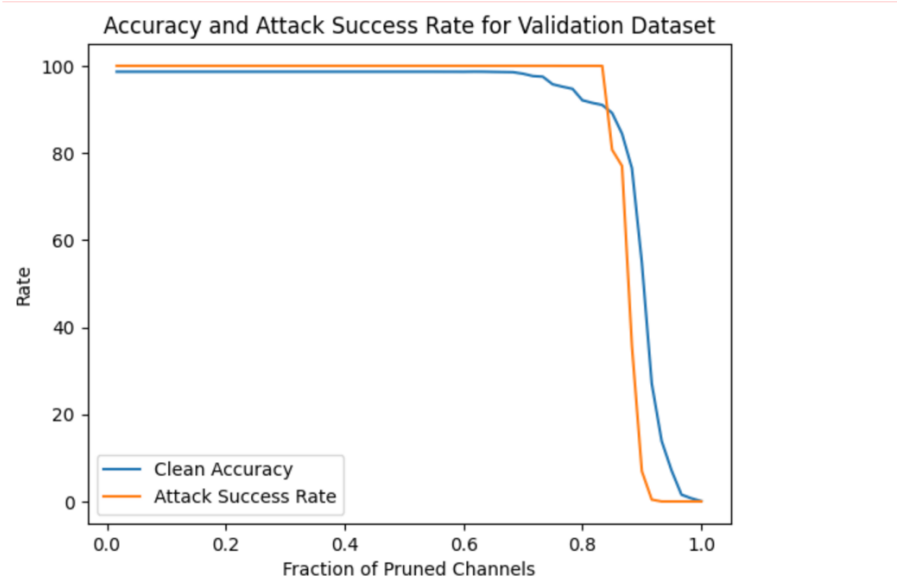
To design a backdoor detector for BadNets trained on the YouTube Face dataset using the pruning defense.

In this lab, the objective is to develop a backdoor detector for BadNets trained on the YouTube Face dataset, employing a pruning defense technique. The first step involves pruning the original BadNet using clean validation data. Channel pruning is performed iteratively, removing one channel at a time based on averaged activations, and the process stops when the accuracy drops by at least X% below the original accuracy.

Subsequently, a GoodNet (G) is constructed by combining the original BadNet and the pruned BadNet. For each test input, both models are used, and if their classification outputs match, the correct class is output; otherwise, it signals a potential backdoor with class N+1.

The evaluation encompasses various scenarios: the original BadNet is tested on a specific backdoor attack (sunglasses backdoor). Repaired networks with pruning percentages of 2%, 4%, and 10% are evaluated using provided scripts, assessing accuracy on clean test data and attack success rate on backdoored test data. Additionally, the GoodNet models are evaluated for accuracy on clean test data and attack success rate on backdoored test data. This comprehensive approach aims to design an effective backdoor detector and evaluate its performance under different scenarios.

Accuracy and Attack Success Rate for Validation Dataset:



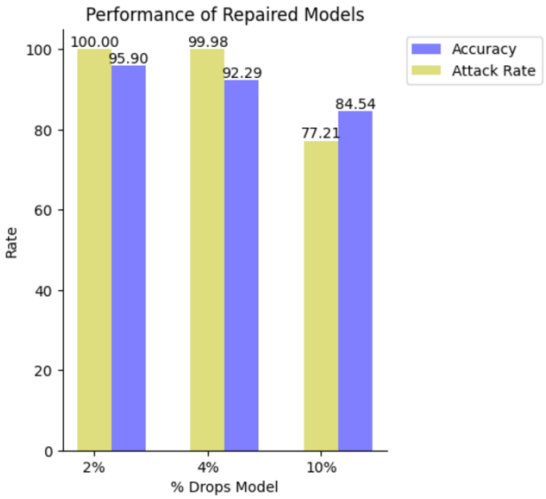
The plots illustrate a significant decline in the success rate of backdoor attacks as a substantial number of neurons are pruned. Initially, the attack success rate remains consistently high at around 100%, with no impact on clean classification accuracy. This initial phase involves pruning neurons that are either zero or poorly activated, rendering them unused by both the honest network and the badnet.

As pruning progresses, specifically when the number of channels removed ranges between 70% and 80% of their initial quantity, a noticeable drop in clean classification accuracy occurs. This indicates the pruning of neurons responsible for classifying clean inputs, while those activated by malicious inputs remain unaffected. Beyond the 80% threshold of neurons removed, both the attack success rate and clean classification accuracy experience a decline, suggesting the removal of neurons crucial for

processing both clean and malicious inputs. In summary, the pruning process demonstrates nuanced effects on backdoor attack success and clean classification accuracy, highlighting the intricate relationship between neuron removal and model behavior.

Evaluation results for repaired models:

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



Evaluation results for GoodNet model:

G_model	G_text_acc	G_attack_rate
G_2%	95.744349	100.000000
G_4%	92.127825	99.984412
G_10%	84.333593	77.209665