Machine Learning For Cyber Security
Lab-4
Ro2151

The models are evaluated based on the number of channels pruned when specific drops in accuracy are observed. The effectiveness of these models is assessed in terms of accuracy on clean test data and attack success rate (ASR) on poisoned (backdoored) test data. Additionally, combined models, named "GoodNet", are evaluated, which integrate predictions from bd_net and a corresponding pruned model.

Methodology
Base Model: Referred to as bd_net.
Pruned Models: Developed by systematically pruning channels from the conv_3 layer of bd_net. Models are saved at points where the accuracy on validation data shows a drop of 2%, 4%, and 10%.
Evaluation Metrics:
Clean Test Accuracy: Accuracy on the dataset not affected by the backdoor attack.
Attack Success Rate (ASR): The rate at which the poisoned samples are misclassified.

The Badnet clean validation accuracy is : 98.64899974019225 %.

Then I pruned the bd_net model before the last pooling layer (conv_3 layer) to obtain a partial model of average activations. However, descending order of average activations did not help at all to yield results so according to the base paper I took ascending order of activations.

Finally by combining all the model_pruned_drop models I created Goodnet models and evaluated them on clean as well as backdoored test inputs with N+1 classes for mismatched predictions (1283).

Results:
Systematic pruning of channels from a specific layer (conv_3) in the model.
Pruning was done incrementally, removing one channel at a time based on specific criteria (such as activation levels).
Key stages were identified where the model's accuracy on a validation dataset dropped by approximately 2%, 4%, and 10% due to pruning.

The first table represents the repaired B’ model whereas the second table represents the performance results of GoodNet on clean test and backdoored input.

Pruning Stage	Total Channels Pruned	Current Accuracy (%)	Attack Success Rate (ASR) (%)
Pre-pruning (Base Model)	0	98.65	100.0
2% Drop	45	95.76	100.0
4% Drop	48	94.34	99.99
10% Drop	52	84.44	77.02

Model Name	Channels Pruned at Drop	Clean Test Accuracy (%)	Attack Success Rate (%)
bd_net	0 (No Pruning)	98.65	100.0
GoodNetdrop2	45	95.62	100.0
GoodNetdrop4	48	94.12	99.99
GoodNetdrop10	52	84.25	77.02