MACHINE LEARNING FOR CYBER SECURITY

BACKDOOR ATTACKS - REPORT

JEEVIKA KANCHERLA (jk7846)

The following assignment has been completed on kaggle

Data Setup:

1. Repository Integration: Imported the CSAW-HackML-2020 repository (https://github.com/csaw-hackml/CSAW-HackML-2020) into the Jupyter notebook for direct model access.
2. Data Gathering: Acquired validation and test datasets, made them public on Google Drive for notebook access.
3. Dataset Details: Used 'bd_valid.h5' and 'bd_test.h5' for validation and testing. These contain images altered with a sunglasses trigger affecting the 'bd_net.h5' model.

Execution Process:

1. Codebase Location: All scripts are in the 'MLcybersec_jk7846.ipynb' file.
2. Data Accessibility: Ensure public access to the data files on Google Drive.
3. Data Importing: Use shared links to obtain file_ids, then download these files into the notebook. Reference screenshot attached.

Approach:

Channel Pruning Strategy: Pruned channels based on average activity in the last pooling layer. Targeted neurons that respond primarily to backdoored inputs. Aimed to minimize impact on accuracy for clean data while reducing effectiveness of backdoor triggers

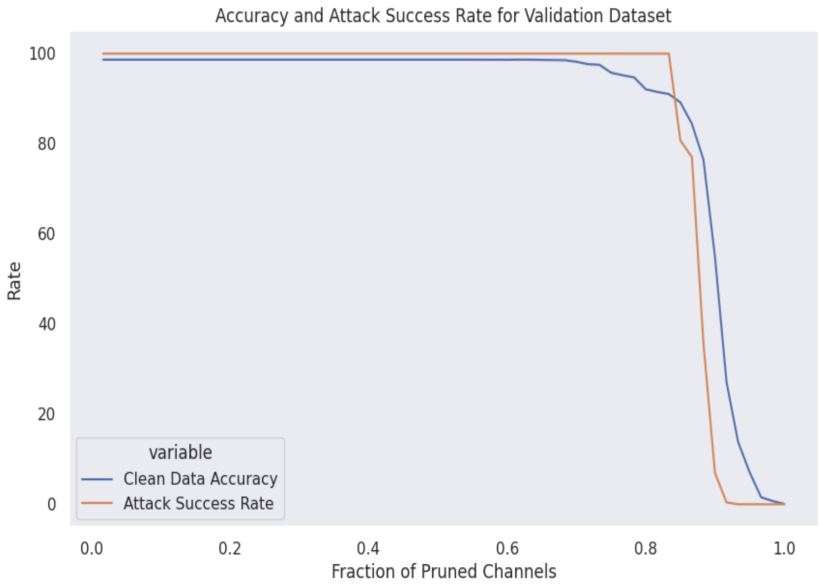
.

Model Saving Criteria: Saved model weights after accuracy drops of 2%, 4%, and 10%. You can find these models in the models folder.

Introduced GoodNet (G) that assesses outputs from both original (B) and pruned (B') BadNet models. G decides the class based on agreement or discrepancy between B and B'.
Findings and Concluding Remarks:

1. **Defense Efficacy**: Pruning defense showed limited success, potentially due to the attack's prune-aware nature.
2. **Neuron Activity and Backdoor Neutralization**: Identified a specific range where backdoor triggers were disabled. Ineffective neurons were pruned, reducing backdoor dataset impact.
3. **Trade-off Concerns:** While attack success rate diminished, this approach also lowered model accuracy.

Observations:

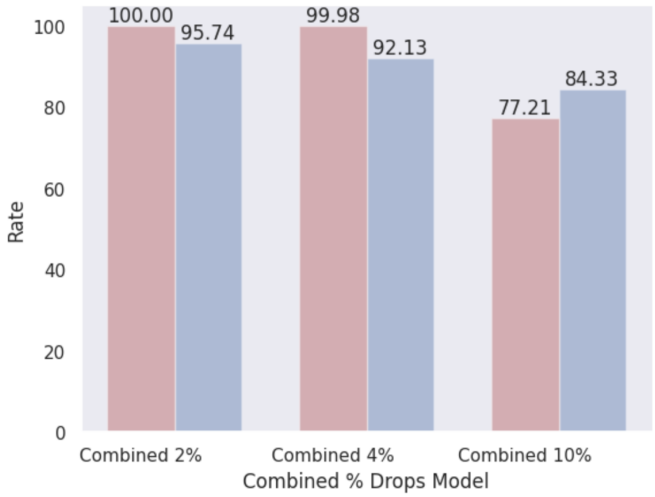


The chart illustrates the variation in two metrics - Clean Data Accuracy and Attack Success Rate - in relation to the proportion of pruned channels in a neural network's validation dataset.

Initially, as channels are pruned, both metrics exhibit high values and show minimal fluctuation, indicating that the network maintains its operational efficiency and security up to a certain level of pruning. However, beyond a specific threshold of pruning, both the accuracy and the attack success rate drop sharply. This sharp downturn at higher levels of channel pruning indicates that the network begins to underperform, losing essential information critical for effective functioning.

The simultaneous decline of both accuracy in classifying correct data and resilience against attacks highlights a strong correlation between the network's classification capability and its security against attacks. This suggests that the network's defense mechanisms are closely tied to its core operational efficacy.

OBSERVATIONS OF COMBINED MODELS:



In the above graph we can notice how the model performs at different rates of drop.

	CLEAN ACCURACY	ATTACK SUCCESS RATE
2% model drop	95.74	100.00
4% model drop	92.13	99.98
10% model drop	84.33	77.21