LAB - 4: Backdoor-Attacks Laboratory Report

By - Sahil Harwani (NetID: sh7253)

Problem Statement Overview

The report addresses countering vulnerabilities in neural networks, specifically focusing on BadNets trained on YouTube Face data. It introduces GoodNet as a solution to detect backdoor inputs while accurately classifying clean data. The YouTube Aligned Face Dataset is utilized for validation and testing.

The defense mechanism involves channel pruning guided by clean validation set metrics, enabling GoodNet to compare outputs between the pruned and original models. Evaluation leverages the DeepID network for facial recognition tasks, using an evaluation script ('eval.py') to assess accuracy with clean data and susceptibility to backdoored inputs.

Observation

- Channel Pruning Impact: Shows a trade-off between increased security (reduced attack success rate) and reduced accuracy on clean data.
- Performance Table: Illustrates accuracy and attack success rate correlation with increased pruning. Results indicate a correlation between increased pruning and decreased attack success rate, though at the cost of reduced accuracy for clean inputs.

Pruned Channels(in fraction)	Clean Data Accuracy(%)	Attack Success Rate(%)
0.02 (2%)	95.900234	100.000000
0.04 (4%)	92.291504	99.984412
0.1 (10%)	84.544037	77.209665

Conclusion

In conclusion, there is a delicate balance between security measures and model performance in neural network design. It sheds light on the potential of pruning strategies to bolster security against backdoor attacks while acknowledging their impact on clean data accuracy. The findings offer insights into implementing defenses in neural network architectures while navigating trade-offs.