LAB - 4: Backdoor-Attacks Laboratory
Shivek Aggarwal (sa7495)

This project addresses the challenge of backdoor vulnerabilities in neural networks by developing a solution to counter compromised classifiers, specifically BadNets trained on the YouTube Face dataset. The resulting model, GoodNet, is engineered to accurately classify clean inputs while also detecting and labeling backdoored inputs by introducing an additional classification category.

The project relies on the YouTube Aligned Face Dataset, dividing it into a validation set ('valid.h5') for pruning validation and a test set ('test.h5') for evaluating the pruned network's effectiveness.

The project leverages the capabilities of the DeepID network for facial recognition tasks. The 'eval.py' script is the chosen tool for calculating the model's accuracy with clean data and its vulnerability to backdoored inputs, providing quantifiable metrics for the network's theoretical robustness.

Channel pruning from the network suggests an increased potential for security by reducing the success rate of attacks. However, this heightened security comes at the cost of a decline in the accuracy of the network with clean data, indicating a trade-off between security and usability.

Observation

A detailed table captures the nuanced outcomes of channel pruning on model performance:

Fraction of Pruned Channels (X)	Clean Data Accuracy (%)	Attack Success Rate (%)
0% (Unpruned)	98.64899974019225	100.0
2%	95.90023382696803	100.0
4%	92.29150428682775	99.98441153546376
10%	84.3335931410756	77.20966484801247

The data illustrates a direct correlation between increased pruning and a reduction in the attack success rate, especially beyond the 10% threshold. This finding suggests the pruning strategy's potential role as a defense against backdoor attacks, albeit with a associated decrease in classification accuracy for clean inputs.

The codebase for this project was developed and executed in Colab Notebook environment on a MacBook with an M1 chip, chosen for its reliability and the efficiency of the M1 chip in handling computational tasks. To replicate the study, users should ensure consistency with the colab Notebook setup, accurate path configurations, and correct execution of the 'eval.py' script with necessary arguments.

Conclusion:

The trade-offs highlighted in this study between security measures and model performance hold critical importance in neural network design. While pruning channels proved effective in reducing the attack success rate, it also impacted the model's accuracy with clean data. The results provide valuable insights into the nuances of machine learning security, emphasizing the need for a careful balance when implementing defenses in neural network architectures.