Backdoors Report

Sayam Dhingra
sd5292@nyu.edu
N16451815

1 Introduction

This report analyzes the impact of channel pruning on a neural network model, focusing on its accuracy with clean test data and its resilience to backdoor attacks. Channel pruning is a strategy used to reduce the complexity of a model by eliminating channels (features) from convolutional layers that contribute less to the output.

2 Methodology

- **Model Used:** A Convolutional Neural Network (CNN) with multiple layers.
- **Datasets:** Clean test data and backdoored test data.
- **Pruning Strategy:** Channels with the lowest average activation in the final pooling layer (`pool_3`) were pruned.

3 Results

The following table summarizes the model's performance on clean test data and backdoored test data, corresponding to different levels of channel pruning.

Fraction(X)	**Accuracy (%)**	**Attack Success Rate (%)**
Original Model (No Pruning)	98.65	100.0
2%	95.90	100.0
4%	92.29	99.98
10%	84.54	77.21

Table 1: Model Performance with Varying Levels of Channel Pruning

The same can be seen in a bar graph where compare the data side-by-side

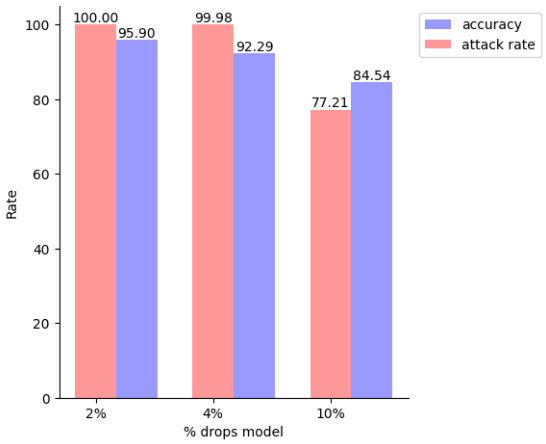

Figure 1: Results

4 Conclusion

The results indicate that increasing the fraction of pruned channels reduces the accuracy on clean test data and the attack success rate on backdoored data. Notably, the model retains high accuracy (above 84%) and significantly lowers the success rate of the attack (to 77.21%) when 10% of the channels are pruned. This demonstrates that channel pruning is effective not only in simplifying the model but also in reducing its vulnerability to backdoor attacks.

5 Limitations and Future Work

- **Model Complexity:** Further analysis is needed to understand the trade-off between model complexity and performance.
- **Diverse Datasets:** Testing on various datasets could provide more comprehensive insights into the pruning strategy's effectiveness.
- **Advanced Pruning Techniques:** Exploring different pruning methods might yield better results in terms of maintaining accuracy while ensuring security against attacks.