Evaluation of Pruned Models as Backdoor Detectors in Neural Networks

Ashwath Shankarnarayan

December 3rd 2023

1 Introduction

This report presents the findings from an experiment designed to evaluate the efficacy of pruned models as backdoor detectors in neural networks, particularly focusing on BadNets. The context of this study is set against the backdrop of increasing concerns regarding the security of neural networks against backdoor attacks.

2 Methodology

The methodology involved pruning the last pooling layer of a pre-trained BadNet model at different levels (2%, 4%, and 10%) and then assessing these models for their accuracy on clean test data and their attack success rate on backdoored test data. The pruning process aimed at identifying the optimal trade-off between model accuracy and security against backdoor threats.

3 Results

The results are summarized in the table below, which presents the accuracy on clean test data and the attack success rate as a function of the fraction of channels pruned.

Fraction of Channels Pruned	Accuracy (%)	Attack Success Rate (%)
2%	95.88	100.0
4%	94.61	99.98
10%	84.46	76.17

Table 1: Model performance as a function of pruning

4 Conclusion

The findings of this study indicate that pruning can be an effective strategy in enhancing the security of neural networks against backdoor attacks. However, it is essential to balance the level of pruning to maintain the overall accuracy of the model. Future

research may explore more sophisticated pruning techniques or alternative methods to safeguard neural networks.