Evaluation of Pruning-Based Backdoor Defense in Neural Networks

* Gunjan Dayani
* gd2275

Backdoor attacks in neural networks are a rising concern as they can cause models to behave incorrectly when triggered by an attacker's input. We explore the effectiveness of a defensive strategy that prunes channels from neural networks to reduce the impact of these backdoors.

**Methodology:**

The defense strategy involved pruning the final pooling layer's channels of a BadNet model in decreasing order of average activation values until the accuracy on a clean validation set dropped by a predefined threshold (X%). Models were saved at thresholds of 2%, 4%, and 10% drops in accuracy. The performance of these pruned models was then compared on clean and poisoned test datasets.

**Results:**

The results are summarized in the following table and figures:

Repaired model

Table 1: Accuracy on clean test data and attack success rate as a function of the fraction of channels pruned.

A screenshot of a graph

Description automatically generated

GoodNet model

Table 2: Accuracy on clean test data and attack success rate as a function of the fraction of channels pruned.

A screenshot of a phone

Description automatically generated

**Discussion:**

The results indicate that pruning can effectively decrease the attack success rate on poisoned data, with a minimal drop in accuracy on clean data. However, as the pruning intensity increased to 10%, the clean data accuracy significantly dropped, which suggests a trade-off between maintaining model utility and mitigating backdoor effects.

**Conclusion:**

Pruning is a viable defense against backdoor attacks in neural networks, capable of reducing the attack success rate while preserving the accuracy on clean data to a considerable extent. Future work should investigate the balance between model utility and security to determine the optimal pruning level.