Homework 4
Backdoor Detection in BadNets: A Pruning Defense

Varuni Buereddy
vb2386

GitHub repo for this project: Here

Backdoor neural networks, or BadNets, pose a significant threat to model security. This report explores the implementation of a pruning defense to detect and mitigate backdoors in BadNets, focusing on the "sunglasses backdoor" scenario.

Defense Model:

- **Data preparation:** The YouTube Face dataset is used for training and validation. The validation set denoted as Dvalid, consists of clean, labeled images, and the BadNet B is subject to a "sunglasses backdoor."
- **Pruning Process:** The last pooling layer of BadNet B is pruned iteratively, removing one channel at a time. Pruning continues until the validation accuracy drops by X%, where X is the defined pruning threshold.
- **GoodNet (G) Construction:** The repaired BadNet (B') resulting from pruning is utilized to construct the GoodNet (G). G has N+1 classes, and its decision rule is based on the agreement between B and B' for a given input.
- **Evaluation:** The defense mechanism is evaluated on clean and backdoored inputs, assessing the accuracy of B, B', and G.

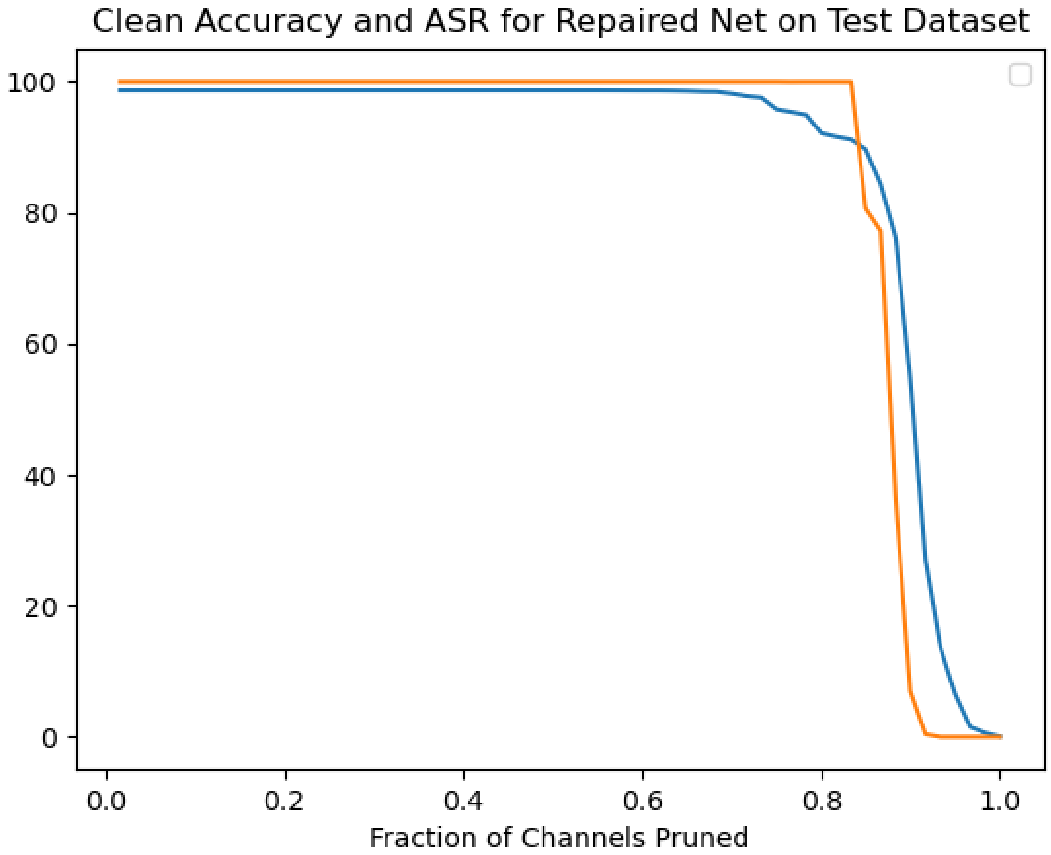
Results and Observations:

Accuracy Metrics

- **Original BadNet (B), Pruned BadNet (B'), GoodNet (G): Clean Accuracy, Backdoor Success Rate**

Pruned Network Performance on Test Set:

Test Dataset	Threshold = 2%	Threshold = 4%	Threshold = 10%
Clean Accuracy	95.90%	92.29%	85.54%
Attack Success Rate	100%	99.98%	77.209%



Observations:

1. **Trade-off between Size and Accuracy:** Pruning Defense is a trade-off between model size and accuracy. By removing certain weights or neurons, the model becomes more compact, but this reduction can result in a loss of information and, consequently, a decrease in accuracy.
2. **Threshold Selection:** The pruning threshold, which determines which connections or nodes to prune, is a critical parameter. Setting it too aggressively may remove important information and lead to bad accuracies, while setting it too conservatively may not result in significant model defense.