Backdoor Lab Report

Aniket Pant - ap7584

Overview:

This lab focuses on developing a defense mechanism against backdoor attacks in neural networks, specifically targeting a BadNet trained on the YouTube Faces dataset. The objective is to implement a pruning strategy that removes potentially compromised channels from the last pooling layer of the network, thereby mitigating the effects of the backdoor without significantly compromising the model's performance on clean data.

Procedure:

Model Architecture: Utilized an established BadNet architecture defined in architecture.py, consisting of convolutional layers followed by max pooling and dense layers, culminating in an output layer designed to classify YouTube Face dataset images.

Pruning Strategy: Adopted a pruning approach that evaluates the average activation of channels in the 'pool_3' layer on a set of clean validation data. Channels were then pruned iteratively, starting with the highest average activation, with the process halting when validation accuracy dropped by a predefined threshold (X%).

Implementation:

- Loaded and preprocessed clean validation data, ensuring the input format matched the model's expectations.
- Defined the calculate_average function to ascertain activation levels within 'pool_3'.
- Implemented prune model function that prunes channels and compiles the pruned model for evaluation.
- Applied a custom deactivate_channel function that deactivates selected channels by zeroing out weights rather than removing them entirely, thus preserving the architecture. (did this because it was getting rather complex if I had to change entire architecture).

Evaluation:

- Compiled the original model with an appropriate optimizer and loss function for performance evaluation.
- Tried to conduct pruning but it is not working, did however move forward just to practice: developed GoodNet (G), a composite model that runs inputs through both the original and pruned models to classify clean inputs or flag backdoored inputs.

Theoretical Conclusion: (couldn't execute but will explain my understanding of the concept)

- **Model Performance**: After pruning, the GoodNet model should effectively classify clean inputs and identify backdoored inputs when the original and pruned models' outputs diverge.
- **Challenges**: Encountered several issues related to data shape compatibility and verbose output during the pruning process. These were resolved by adjusting the data transpose operations and setting verbose flags appropriately.
- **Effect of Threshold:** While a lower threshold preserves accuracy, it may be insufficient for backdoor mitigation. Conversely, a higher threshold more effectively counters the backdoor but risks degrading the model's performance.

In conclusion, the lab demonstrates a viable strategy for mitigating backdoor threats in neural networks through targeted pruning.