ECE-GY-9163 LAB 4: Backdoor detector for BadNets

Pratham Mehta
pm3483

Aim:
In this lab, I applied a pruning strategy to a BadNet model that had been trained using the YouTube Face dataset. The technique involves eliminating neurons from the convolutional layer located immediately before the final pooling layer. To determine which neurons to prune, I calculated the feature map for the last pooling layer, average these values to obtain the activation levels, and then organize them in ascending order. Then I had to proceed to prune neurons based on their activation values until the accuracy level drops to the specified threshold. I had conducted this process using three distinct accuracy thresholds: 2%, 4%, and 10%.

Results:

Threshold	Channel Pruned	Clean Accuracy	Attack Success Rate
2	75%	95.90	100.00
4	80%	92.29	99.98
10	86.7%	84.54	77.21

Observing that a decrease of 2% in performance occurred only after pruning 75% of the neurons suggests that the majority of neurons were redundant. This finding is in agreement with the concepts previously explored in our class discussions.

Github Link:
https://github.com/Pratham-mehta/backdoor-detector-for-BadNets