ECE-GY-9163 HW2: Backdoor detector for BadNets

Akash Mishra am11533

Introduction

In this HW, we are implementing pruning defense on a BadNet trained on YouTube Face dataset. It is done by pruning neurons from the convolution layer just prior to last pooling layer. We get the feature map for the last pooling layer, average it to get the activation value which is again sorted in the increasing order. Using this ascending activation value for channel, we start pruning till we reach the provided threshold for accuracy. For this work, we had three thresholds: 2,4 and 10%.

Results

Threshold	Channel Pruned	Clean Accuracy	Attack Success Rate
2	75%	95.90	100.0
4	80%	92.29	99.98
10	86.7%	84.54	77.21

Seeing as 2% drop was seen only after 75% of the neurons were pruned indicating that most of the neurons were not useful which aligns with what has been previously discussed in the class also.

GitHub Link: https://github.com/akashsky1994/backdoor-detector-badnets