ML4SEC HW2: Backdoor detector for BadNets

Luca Collini (lc4976)

1. Introduction

This homework consisted in repairing a backdoored neural network using the pruning defense technique. The defense mechanism works by pruning a neuron at a time from the last pooling layer. The neurons are pruned in order from lowest to highest average activation value on the clean validation dataset. The pruning is performed until the accuracy on the validation dataset is corrupted more than a set threshold. For this homework we had 3 thresholds: 2,4, and 10%.

2. Results

The following table reports the accuracy on clean test data and the attack success rate (on backdoored test data).

Threshold	Ch. pruned	Accuracy	Attack S.R.
2%	73%	95.74%	100%
4%	78%	92.12%	99.98%
10%	85%	84.33%	77.2%

We can guess that a lot of neurons are not useful for the computation as pruning 73% of the neurons in the last pooling layer lowers accuracy of only 2 percent.