ML for Cybesec – Homework Lab 02 – Backdoor Defenses
Net Id: vg2097

The following lab explains how easy it is to do a backdoor attack on a regular model and how a model with very high classification accuracy can also be model which is highly vulnerable and can get the attack success rate very high model.

We need to take certain measures to make the model robust to these attacks to a certain attack.
Here we iteratively prune the model layers to remove the possible backdoor activations in the model resulting in reducing the backdoor activation neurons and in-turn increasing the model robustness against such attacks. However with the following pruning process as we increasingly remove/turn off certain neurons the model accuracy in classifying the test data also decreases.

We then decide based on our requirement to select the trade-off between the model accuracy Vs Attack success rate on the model.
Below is the table for the model's accuracy and the attack success rate and the amount of pruning.

	prune index	clean test accuracy	Attack success Rate
0	0.016666666666666700	98.64899974019230	100.0
1	0.03333333333333330	98.64899974019230	100.0
2	0.05	98.64899974019230	100.0
3	0.06666666666666670	98.64899974019230	100.0
4	0.08333333333333330	98.64899974019230	100.0
5	0.1	98.64899974019230	100.0
6	0.11666666666666700	98.64899974019230	100.0
7	0.13333333333333300	98.64899974019230	100.0
8	0.15	98.64899974019230	100.0
9	0.16666666666666700	98.64899974019230	100.0
10	0.18333333333333300	98.64899974019230	100.0
11	0.2	98.64899974019230	100.0
12	0.21666666666666700	98.64899974019230	100.0
13	0.23333333333333300	98.64899974019230	100.0
14	0.25	98.64899974019230	100.0
15	0.26666666666666700	98.64899974019230	100.0
16	0.2833333333333330	98.64899974019230	100.0
17	0.3	98.64899974019230	100.0
18	0.31666666666666700	98.64899974019230	100.0
19	0.3333333333333330	98.64899974019230	100.0

20	0.35	98.64899974019230	100.0
21	0.36666666666666700	98.64899974019230	100.0
22	0.38333333333333300	98.64899974019230	100.0
23	0.4	98.64899974019230	100.0
24	0.4166666666666670	98.64899974019230	100.0
25	0.43333333333333300	98.64899974019230	100.0
26	0.45	98.64899974019230	100.0
27	0.4666666666666670	98.64899974019230	100.0
28	0.48333333333333300	98.64899974019230	100.0
29	0.5	98.64899974019230	100.0
30	0.5166666666666670	98.64899974019230	100.0
31	0.5333333333333330	98.64899974019230	100.0
32	0.55	98.64899974019230	100.0
33	0.5666666666666670	98.64899974019230	100.0
34	0.5833333333333330	98.64033948211660	100.0
35	0.6	98.63167922404090	100.0
36	0.6166666666666670	98.65765999826800	100.0
37	0.6333333333333330	98.64899974019230	100.0
38	0.65	98.6056984498138	100.0
39	0.6666666666666670	98.57105741751100	100.0
40	0.6833333333333330	98.53641638520830	100.0
41	0.7	98.19000606218070	100.0
42	0.7166666666666670	97.65307006148780	100.0
43	0.7333333333333330	97.50584567420110	100.0
44	0.75	95.75647354291160	100.0
45	0.7666666666666670	95.20221702606740	99.97661730319560
46	0.7833333333333330	94.7172425738287	99.98441153546380
47	0.8	92.09318437689440	99.98441153546380
48	0.8166666666666670	91.49562656967180	99.98441153546380
49	0.8333333333333330	91.01931237550880	99.97661730319560
50	0.85	89.17467740538670	80.6469212782541
51	0.8666666666666670	84.43751623798390	77.20966484801250
52	0.8833333333333330	76.48739932449990	36.26656274356980
53	0.9	54.8627349095003	6.96024941543258
54	0.9166666666666670	27.08928726076040	0.4208885424785660
55	0.9333333333333330	13.87373343725640	0.0
56	0.95	7.101411622066340	0.0
57	0.9666666666666670	1.5501861955486300	0.0
58	0.9833333333333330	0.7188014202823240	0.0
59	1.0	0.0779423226812159	0.0