CS-89.31: Deep Learning Generalization and Robustness

Amittai Siavava

05/23/2023

1. Adversarial Training

Adversarial training took about 2 hours on my laptop (which has a quite capable GPU). The general trend was improvement in both the benign and adversarial test accuracies the more the model was trained. However, the rate of improvement slowed down and became almost zero, suggesting that the methods used would reach a limit and perhaps other methods would be needed to improve the model further.

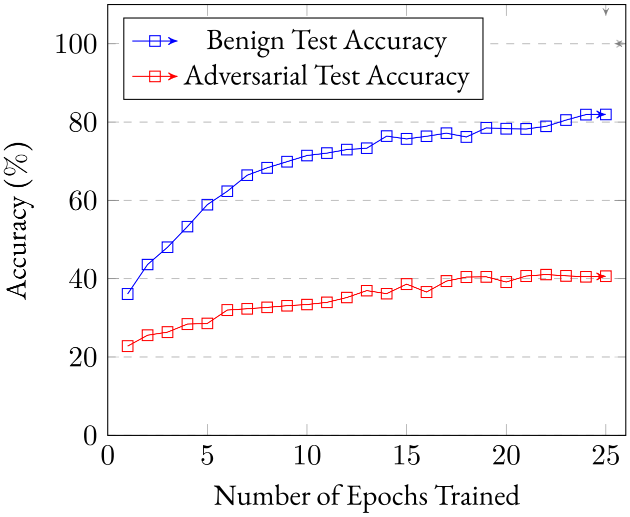

Figure 1. Adversarial Training: Model Performance vs. Epochs Trained

Number of Epochs	Benign Test Accuracy	Adversarial Test Accuracy
1	36.10	22.78
2	43.62	25.56
3	48.02	26.35
4	53.31	28.41
5	58.91	28.57
6	62.32	31.99
7	66.42	32.32
8	68.31	32.67
9	69.88	33.10
10	71.45	33.40
11	72.06	33.95
12	72.97	35.19
13	73.31	36.95
14	76.40	36.18
15	75.69	38.61
16	76.36	36.59
17	77.15	39.38
18	76.19	40.42
19	78.51	40.48
20	78.30	39.15
21	78.23	40.68
22	78.90	41.06
23	80.50	40.73
24	81.90	40.50
25	81.94	40.62

Table 1. Adversarial Training: Model Performance vs. Epochs Trained

2. Data Augmentation

In the data augmentation part, my results seemed to improve more with more training on augmented data. I think this is because the model was able to generalize better after having seen samples of the data in different forms, i.e. flipped, rotated, cropped, and with hue changes.

Mode	10 Epochs	30 Epochs	50 Epochs
Tech0	0.41409090161323455	0.65854935935839753	0.7345454575483436
Tech1	0.44508402584390405	0.68590909123420715	0.7795454421496957
Tech2	0.43699999910593033	0.73999999701976776	0.7995454454421997
Tech3	0.47045454859733582	0.75090909540653623	0.8457935293577323

Table 2. Training Results with Data Augmentation

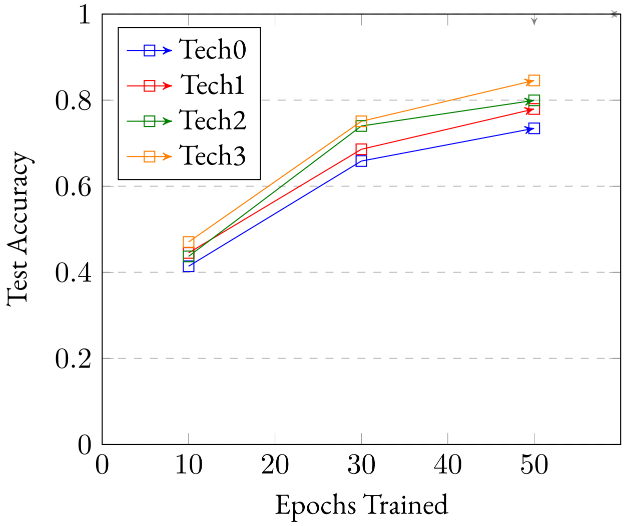

Figure 2. Data Augmentation: Model Performance vs. Epochs Trained