CS-89.31: Deep Learning Generalization and Robustness

Amittai Siavava

05/23/2023

1. Adversarial Training

The adversarial training part was quite interesting. I initially tried training on a MacBook Pro but it was taking too long (around 5 minutes per epoch). I then shifted to my Windows laptop which has a discrete graphics card, and the training time went down to less than half a minute per epoch. The entire training still took around 90 minutes, but it was really interesting to see the model predictions improve.

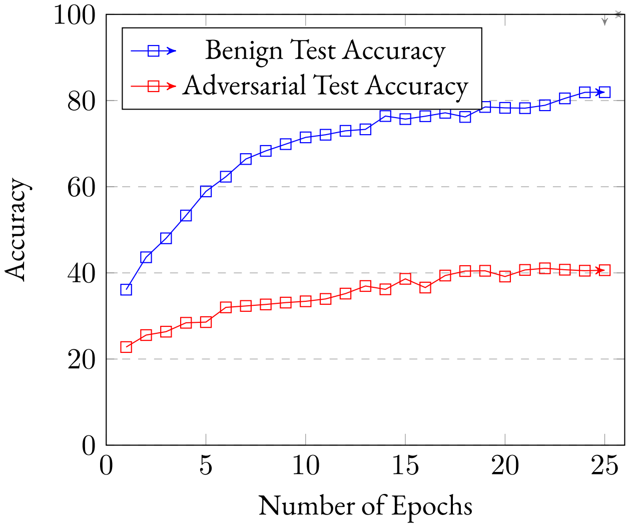

Figure 1. Adversarial Training Results (Graph).

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 Results.

2. Data Augmentation

In the data augmentation part, with some particular methods, the model seemed to perform worse the more I trained it.