CS579 – Homework 3 Solutions

Opeyemi Ajibuwa

Task 1

a. The plot of the {ratio of poisons in the training set} vs. {classification accuracy} on the test-set.

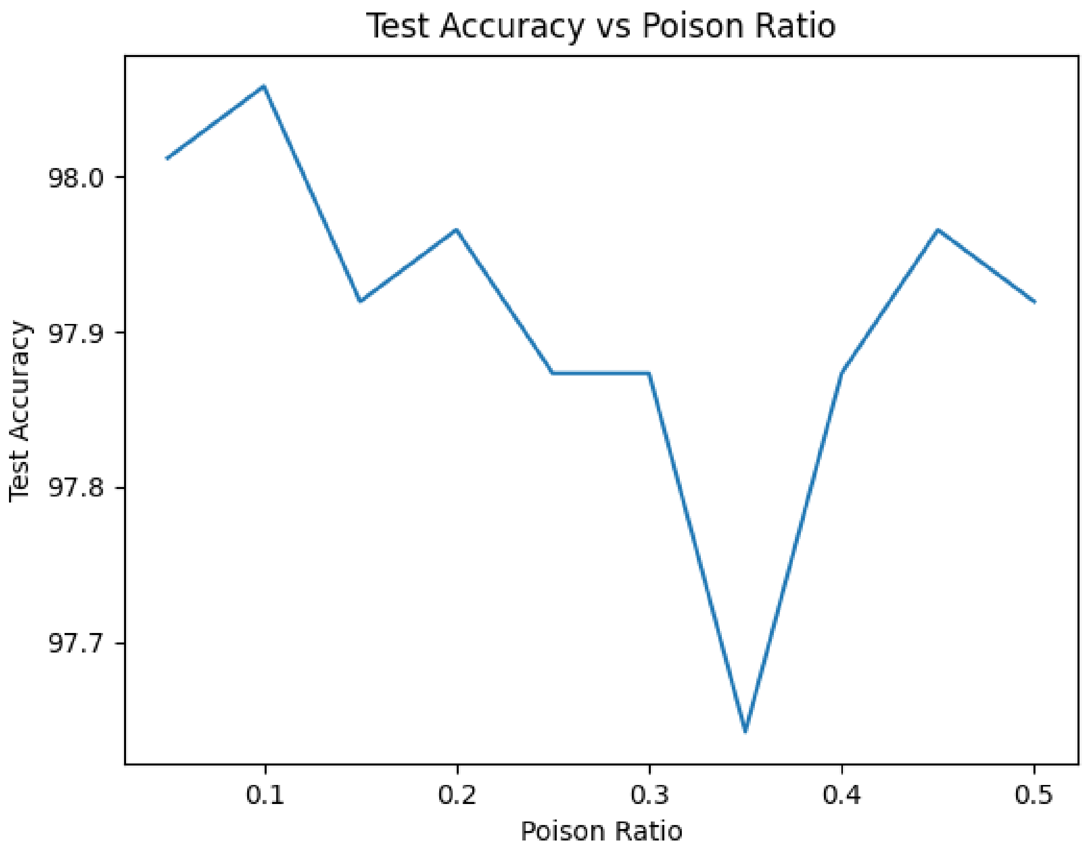

Figure 1

Results analysis: The results obtained under this attack differ significantly from expectation. I expected that the test accuracy will decrease as the ratio of poisoned examples in the training dataset increases. However, this was not quite the case. As seen in Fig. 1 above, there was marginal degradation in the test accuracy as the percentage of poisoned examples is increased in the training dataset. The test accuracy is highest when the poison ratio is 0.1, which is better than with a clean test set (i.e., when poison ratio = 0). The test accuracy further decreases with the lowest degradation at poison ratio = 0.35 after which the test accuracy starts to increase again until the last test case. A possible explanation for this odd observation may be that the implemented random label flipping attack is weak or that the trained logistic model is robust against the generated poisoned outlier.

Task II

Number of poisons	1	5	10	25	50	100
Number of successful attacks over the 5 targets	0/5	2/5	4/5	5/5	5/5	5/5

Table 1

Results analysis: As the table above shows, the number of successful targeted poisoning attack on the ResNet model increases as the number of poisons introduced into the training dataset increases. This observation aligns well with expectations. Over all the five targets, no misclassification is achieved by introducing a single poisoned base example. However, the attack becomes more successful as larger poisons samples are introduced.