Adversarial Learning - Assignment 5 Eitamar Saraf, Moshe Nasletashvili

1. detector over those 400 adversarial examples

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FGSM | TGSM | PGD | Targeted PGD | Total |
| Correctly classified | 1.0 | 1.0 | 0.99 | 0.99 | 0.995 |
| Correctly identified | 0.97 | 0.96 | 0.58 | 0.48 | 0.7475 |
| Wrongly classified | 0 | 0 | 0.01 | 0.01 | 0.005 |

1. detector over 400 randomly selected VALID inputs
   1. False positive rate: 0.535
2. Parameters:
   1. FGSM: epsilon=0.3
   2. TGSM: epsilon=0.3
   3. PDG & Targeted PGD: epsilon=3, iter\_eps = 0.03, iterations=60, min\_x=-1.0, max\_x=1.0
   4. Detector: epsilon: 0.1, iteration: 1000, lr = 0.5, lr\_decay after 500 iterations to 0.05