Part 1

As a baseline, the base, defenseless model was evaluated on benign and adversarial images. The model was evaluated on clean and adversarial images with three input image transformation defenses: JPEG compression, resizing, and blur. The recorded metrics for each test were accuracy on clean images (1), accuracy on adversarial images (2), and success rate of adversarial images (3):

The results of the baseline and defense tests are displayed in table 1.

Table 1. Accuracy scores per by defense with naïve PGD

|  |  |  |  |
| --- | --- | --- | --- |
| Defense | Benign Accuracy (1) | AE Accuracy (2) | Attack Success (3) |
| None | 77.00% | 5.00% | 93.00% |
| JPEG Compression | 72.00% | 33.00% | 57.00% |
| Resizing | 73.00% | 49.00% | 38.00% |
| Blurring | 73.00% | 64.00% | 22.00% |

The results show increases in AE accuracy and attack success percentages with slight decreases in benign accuracy between the baseline and the three defenses. More specifically, blurring is the best defense, raising the AE accuracy by 59.00% and decreasing attack success by 71.00%, and only decreasing benign accuracy by 4.00%. Image resizing performed the second best, raising AE accuracy by 44.00% and decreasing attack success by 55.00% with a benign accuracy decrease of 4.00%. Of the three defenses, JPEG compression performed the worst, increasing AE accuracy by only 28.00%, decreasing attack success by 36.00%, and decreasing benign accuracy by a slightly higher 5.00%.

Part 2

To break the transformation defense created in Part 1, an expectation over transformation attack was created, averaging the gradients across each of the three transformations. The attack was evaluated on the four defense cases used for evaluating Part 1 with the three accuracy scores.

Table 2. Accuracy scores per defense with EOT

|  |  |  |  |
| --- | --- | --- | --- |
| Defense | Benign Accuracy (1) | AE Accuracy (2) | Attack Success (3) |
| None | 77.00% | 23.00% | 71.00% |
| JPEG Compression | 72.00% | 24.00% | 71.00% |
| Resizing | 73.00% | 15.00% | 80.00% |
| Blurring | 73.00% | 9.00% | 86.00% |

Interestingly, the EOT attack performed best on the defenses that performed best in Part 1 with an 86.00% success rate with the blurring transformation and an 80.00% success rate with resizing, compared to a 71% success rate on the JPEG transformation and the defenseless model. This indicates that the attack is overfitting its images towards the more potent defenses, perhaps because their labels are more confident and influence the gradient more when the average is taken across the three defenses.