

Supplementary Material: Catastrophic Child’s Play: Easy to Perform, Hard to Defend Adversarial Attacks

Chih-Hui Ho^{1*} Brandon Leung^{1*} Erik Sandström² Yen Chang¹ Nuno Vasconcelos¹
¹University of California, San Diego ²Lund University
`{chh279, b7leung}@ucsd.edu` `tfy14esa@student.lu.se` `{yec084, nvasconcelos}@ucsd.edu`

A. Amazon Turk test times

One of the variables in the experimental protocol used to measure perturbation perceptibility is the time for which images are shown to the subjects. Several works have shown that neural activity exhibits signs of object recognition within about 200 ms of an image stimulus, with reaction times about 150 ms later. Preliminary experiments with a 350 ms viewing time showed that this was too little, at least for Turk experiments. Turkers only identified the TP as being the same image 55 percent of the time. While they did much better at rejecting different objects, this time was considered overall too aggressive. Subsequent experiments with a longer limit of 750 ms suggested that this was enough time. The IPRs obtained with the two settings are shown in Table 1.

| | p^{TP} | p^{CS} | p^{PV} | p^{DO} |
|--------|----------|----------|----------|----------|
| 350 ms | 0.551 | 0.404 | 0.258 | 0.059 |
| 750 ms | 0.977 | 0.798 | 0.106 | 0.010 |

Table 1: Preliminary Amazon Turk A/B testing results, wherein turkers were given 350 ms or 750 ms to remember images. The turkers’ average imperceptibility scores in the context of image recognition reveal similar trends relative to their respective upper bound p^{TP} and lower bound p^{DO} . However, elements such as fatigue may play a factor and thus 750 ms was ultimately chosen in the experiment design.

B. Recognition rates (RR) for IP and SIP

Tables 2-3 summarize the RRs of IP and SIP attacks per model, defense dataset, and defense algorithm. While, in general, ResNet outperformed the other models, the effect of the attacks on the three models was quantitatively similar. Defense algorithms were more effective for IPs than SIPs.

This is not surprising, since the former tend to be smaller perturbations. The largest gains were obtained by using defense datasets augmented with CS and PV perturbations (‘All’).

C. Example adversarial samples

Additional adversarial samples of CS-IP, PV-IP, CS-SIP and PV-SIP are provided in Table 4, 5, 6 and 7 respectivly.

*Equal contribution

Table 2: Recognition rate (RR) for IP.

| | | ImageNet | | | | Frontal | | | | All | | | |
|-----------------------|--------------|----------|--------|------|------|---------|--------|------|------|------|--------|------|------|
| Defense | | Alex | ResNet | VGG | Avg | Alex | ResNet | VGG | Avg | Alex | ResNet | VGG | Avg |
| Camera shake attack | | | | | | | | | | | | | |
| Aug. | None | 76.0 | 78.5 | 69.2 | 74.6 | 82.3 | 88.2 | 88.5 | 86.3 | 83.0 | 92.3 | 92.5 | 89.3 |
| | Affine | 77.8 | 76.6 | 74.5 | 76.3 | 85.4 | 86.2 | 89.8 | 87.1 | 84.0 | 93.8 | 88.0 | 88.6 |
| | Blur | 76.5 | 80.1 | 72.5 | 76.4 | 83.0 | 84.2 | 90.5 | 85.9 | 81.4 | 94.9 | 89.8 | 88.7 |
| | Blur-Affine | 78.0 | 72.8 | 76.1 | 75.7 | 86.8 | 86.0 | 87.2 | 86.7 | 83.7 | 91.2 | 88.2 | 87.7 |
| | Worst | 68.0 | 77.2 | 72.0 | 72.4 | 88.7 | 88.8 | 88.5 | 88.7 | 84.0 | 91.8 | 90.0 | 88.6 |
| | Color Jitter | 77.3 | 78.5 | 70.3 | 75.4 | 87.4 | 90.4 | 91.9 | 89.9 | 89.9 | 92.6 | 88.4 | 90.3 |
| Adv. | Avg | 75.5 | 77.1 | 73.1 | 75.2 | 86.3 | 87.1 | 89.6 | 87.7 | 84.6 | 92.9 | 88.9 | 88.8 |
| | FGSM | 76.1 | 83.0 | 70.7 | 76.6 | 84.3 | 90.9 | 84.5 | 86.6 | 86.1 | 84.8 | 91.0 | 87.3 |
| | ENS | 74.1 | 82.0 | 78.2 | 78.1 | 87.6 | 83.7 | 86.7 | 86.0 | 82.5 | 81.2 | 89.6 | 84.4 |
| | IFGSM | 70.7 | 77.1 | 73.6 | 73.8 | 85.1 | 88.1 | 88.3 | 87.2 | 82.8 | 86.7 | 88.0 | 85.8 |
| Pose variation attack | | | | | | | | | | | | | |
| Aug. | None | 79.5 | 81.1 | 72.2 | 77.6 | 80.6 | 79.7 | 80.9 | 80.4 | 78.3 | 91.8 | 84.5 | 84.9 |
| | Affine | 62.2 | 83.0 | 54.5 | 66.6 | 89.5 | 67.8 | 81.0 | 79.4 | 83.1 | 88.7 | 85.9 | 85.9 |
| | Blur | 78.4 | 85.5 | 63.8 | 75.9 | 80.0 | 77.4 | 75.4 | 77.6 | 83.6 | 91.9 | 83.3 | 86.3 |
| | Blur-Affine | 71.8 | 80.4 | 61.7 | 71.3 | 70.0 | 83.6 | 80.0 | 77.9 | 87.7 | 81.9 | 86.8 | 85.5 |
| | Worst | 56.8 | 84.2 | 65.3 | 68.8 | 85.2 | 86.2 | 81.8 | 84.4 | 81.4 | 86.1 | 77.6 | 81.7 |
| | Color Jitter | 78.9 | 88.9 | 73.8 | 80.5 | 79.3 | 85.5 | 87.3 | 84.0 | 84.4 | 94.5 | 88.9 | 89.3 |
| Adv. | Avg | 69.6 | 84.4 | 63.8 | 72.6 | 80.8 | 80.1 | 81.1 | 80.7 | 84.0 | 88.6 | 84.5 | 85.7 |
| | FGSM | 83.8 | 90.7 | 57.8 | 77.4 | 83.6 | 82.3 | 84.1 | 83.3 | 80.7 | 83.1 | 83.3 | 82.4 |
| | ENS | 66.7 | 78.2 | 57.9 | 67.6 | 88.9 | 79.7 | 83.8 | 84.1 | 72.7 | 83.3 | 78.3 | 78.1 |
| | IFGSM | 34.4 | 71.8 | 60.0 | 55.4 | 72.2 | 76.9 | 75.0 | 74.7 | 85.7 | 88.1 | 78.5 | 84.1 |
| Pose variation attack | | | | | | | | | | | | | |
| Aug. | None | 78.3 | 72.5 | 60.3 | 73.7 | 77.5 | 84.4 | 84.0 | 82.0 | 83.5 | 90.3 | 87.5 | 87.1 |
| | Affine | 68.7 | 75.5 | 71.4 | 71.8 | 80.4 | 83.1 | 86.7 | 83.4 | 82.1 | 89.0 | 84.5 | 85.2 |
| | Blur | 76.2 | 79.6 | 66.8 | 74.2 | 81.9 | 85.5 | 86.9 | 84.8 | 82.7 | 90.7 | 87.2 | 86.9 |
| | Blur-Affine | 79.2 | 72.4 | 74.4 | 75.4 | 81.1 | 85.1 | 84.4 | 83.5 | 84.5 | 90.4 | 89.0 | 88.0 |
| | Worst | 70.0 | 75.8 | 73.3 | 73.0 | 84.3 | 84.1 | 82.8 | 83.8 | 80.8 | 90.4 | 88.1 | 86.4 |
| | Color Jitter | 79.5 | 73.3 | 70.9 | 74.5 | 84.0 | 87.3 | 87.8 | 86.4 | 84.3 | 91.2 | 85.9 | 87.1 |
| Adv. | Avg | 74.7 | 75.3 | 71.4 | 73.8 | 82.3 | 85.0 | 85.7 | 84.4 | 82.9 | 90.3 | 87.0 | 86.7 |
| | FGSM | 75.4 | 77.1 | 66.2 | 72.9 | 81.4 | 87.1 | 85.5 | 84.7 | 79.6 | 81.7 | 88.2 | 83.2 |
| | ENS | 74.0 | 80.9 | 72.2 | 75.7 | 82.9 | 82.2 | 85.7 | 83.6 | 79.5 | 80.2 | 85.9 | 81.9 |
| | IFGSM | 66.2 | 75.4 | 73.8 | 71.8 | 78.1 | 85.1 | 85.3 | 82.8 | 81.4 | 84.5 | 83.9 | 83.3 |
| Pose variation attack | | | | | | | | | | | | | |
| Trans. | None | 40.0 | 52.3 | 49.2 | 47.2 | 58.4 | 67.9 | 64.8 | 63.7 | 72.1 | 82.8 | 82.5 | 79.1 |
| | Affine | 36.5 | 52.2 | 46.7 | 45.1 | 53.3 | 61.1 | 62.1 | 58.8 | 68.5 | 81.6 | 79.6 | 76.5 |
| | Blur | 36.4 | 56.1 | 43.2 | 45.2 | 58.3 | 67.9 | 65.9 | 64.1 | 72.4 | 83.6 | 78.9 | 78.3 |
| | Blur-Affine | 41.5 | 54.7 | 46.4 | 47.5 | 53.0 | 65.3 | 61.8 | 60.0 | 69.1 | 80.1 | 80.4 | 76.6 |
| | Worst | 40.6 | 53.8 | 46.7 | 47.1 | 58.1 | 68.1 | 62.9 | 63.0 | 67.0 | 81.7 | 79.6 | 76.1 |
| | Color Jitter | 39.7 | 50.6 | 46.3 | 45.5 | 52.5 | 67.3 | 65.1 | 61.6 | 73.1 | 83.7 | 80.5 | 79.1 |
| Adv. | Avg | 38.9 | 53.5 | 45.9 | 46.1 | 55.0 | 65.9 | 63.6 | 61.5 | 70.0 | 82.1 | 79.8 | 77.3 |
| | FGSM | 41.3 | 59.5 | 46.8 | 49.2 | 55.4 | 65.4 | 62.6 | 61.1 | 70.5 | 72.8 | 79.6 | 74.3 |
| | ENS | 41.5 | 52.4 | 45.0 | 46.3 | 59.2 | 54.5 | 60.7 | 58.1 | 71.3 | 71.0 | 76.0 | 72.8 |
| | IFGSM | 47.6 | 49.9 | 43.4 | 47.0 | 54.3 | 51.6 | 60.6 | 55.5 | 66.7 | 68.8 | 74.4 | 70.0 |

Table 3: Recognition rate (RR) for SIP

| | | ImageNet | | | | Frontal | | | | All | | | |
|-----------------------|--------------|----------|--------|------|------|---------|--------|------|------|------|--------|------|------|
| Defense | | Alex | ResNet | VGG | Avg | Alex | ResNet | VGG | Avg | Alex | ResNet | VGG | Avg |
| Camera shake attack | | | | | | | | | | | | | |
| Aug. | None | 78.3 | 72.5 | 60.3 | 73.7 | 77.5 | 84.4 | 84.0 | 82.0 | 83.5 | 90.3 | 87.5 | 87.1 |
| | Affine | 68.7 | 75.5 | 71.4 | 71.8 | 80.4 | 83.1 | 86.7 | 83.4 | 82.1 | 89.0 | 84.5 | 85.2 |
| | Blur | 76.2 | 79.6 | 66.8 | 74.2 | 81.9 | 85.5 | 86.9 | 84.8 | 82.7 | 90.7 | 87.2 | 86.9 |
| | Blur-Affine | 79.2 | 72.4 | 74.4 | 75.4 | 81.1 | 85.1 | 84.4 | 83.5 | 84.5 | 90.4 | 89.0 | 88.0 |
| | Worst | 70.0 | 75.8 | 73.3 | 73.0 | 84.3 | 84.1 | 82.8 | 83.8 | 80.8 | 90.4 | 88.1 | 86.4 |
| | Color Jitter | 79.5 | 73.3 | 70.9 | 74.5 | 84.0 | 87.3 | 87.8 | 86.4 | 84.3 | 91.2 | 85.9 | 87.1 |
| Adv. | Avg | 74.7 | 75.3 | 71.4 | 73.8 | 82.3 | 85.0 | 85.7 | 84.4 | 82.9 | 90.3 | 87.0 | 86.7 |
| | FGSM | 75.4 | 77.1 | 66.2 | 72.9 | 81.4 | 87.1 | 85.5 | 84.7 | 79.6 | 81.7 | 88.2 | 83.2 |
| | ENS | 74.0 | 80.9 | 72.2 | 75.7 | 82.9 | 82.2 | 85.7 | 83.6 | 79.5 | 80.2 | 85.9 | 81.9 |
| | IFGSM | 66.2 | 75.4 | 73.8 | 71.8 | 78.1 | 85.1 | 85.3 | 82.8 | 81.4 | 84.5 | 83.9 | 83.3 |
| Pose variation attack | | | | | | | | | | | | | |
| Trans. | None | 40.0 | 52.3 | 49.2 | 47.2 | 58.4 | 67.9 | 64.8 | 63.7 | 72.1 | 82.8 | 82.5 | 79.1 |
| | Affine | 36.5 | 52.2 | 46.7 | 45.1 | 53.3 | 61.1 | 62.1 | 58.8 | 68.5 | 81.6 | 79.6 | 76.5 |
| | Blur | 36.4 | 56.1 | 43.2 | 45.2 | 58.3 | 67.9 | 65.9 | 64.1 | 72.4 | 83.6 | 78.9 | 78.3 |
| | Blur-Affine | 41.5 | 54.7 | 46.4 | 47.5 | 53.0 | 65.3 | 61.8 | 60.0 | 69.1 | 80.1 | 80.4 | 76.6 |
| | Worst | 40.6 | 53.8 | 46.7 | 47.1 | 58.1 | 68.1 | 62.9 | 63.0 | 67.0 | 81.7 | 79.6 | 76.1 |
| | Color Jitter | 39.7 | 50.6 | 46.3 | 45.5 | 52.5 | 67.3 | 65.1 | 61.6 | 73.1 | 83.7 | 80.5 | 79.1 |
| Adv. | Avg | 38.9 | 53.5 | 45.9 | 46.1 | 55.0 | 65.9 | 63.6 | 61.5 | 70.0 | 82.1 | 79.8 | 77.3 |
| | FGSM | 41.3 | 59.5 | 46.8 | 49.2 | 55.4 | 65.4 | 62.6 | 61.1 | 70.5 | 72.8 | 79.6 | 74.3 |
| | ENS | 41.5 | 52.4 | 45.0 | 46.3 | 59.2 | 54.5 | 60.7 | 58.1 | 71.3 | 71.0 | 76.0 | 72.8 |
| | IFGSM | 47.6 | 49.9 | 43.4 | 47.0 | 54.3 | 51.6 | 60.6 | 55.5 | 66.7 | 68.8 | 74.4 | 70.0 |

Table 4: Adversarial samples for CS IP

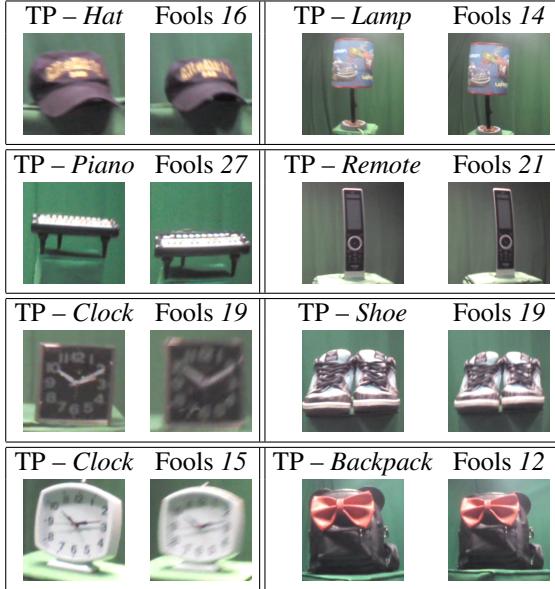


Table 5: Adversarial samples for PV IP

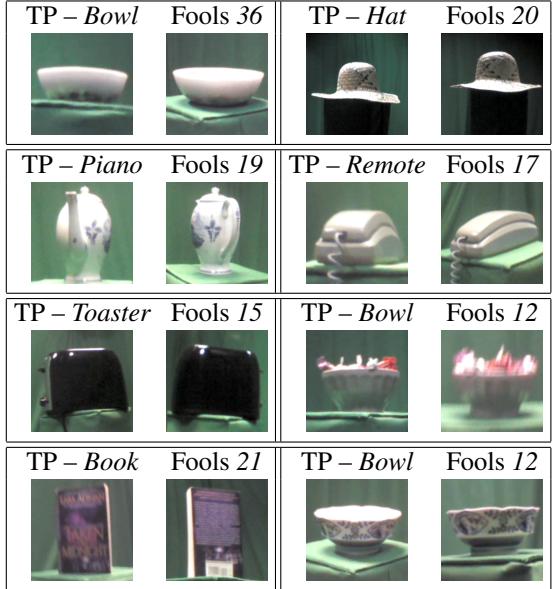


Table 6: Adversarial samples for CS SIP

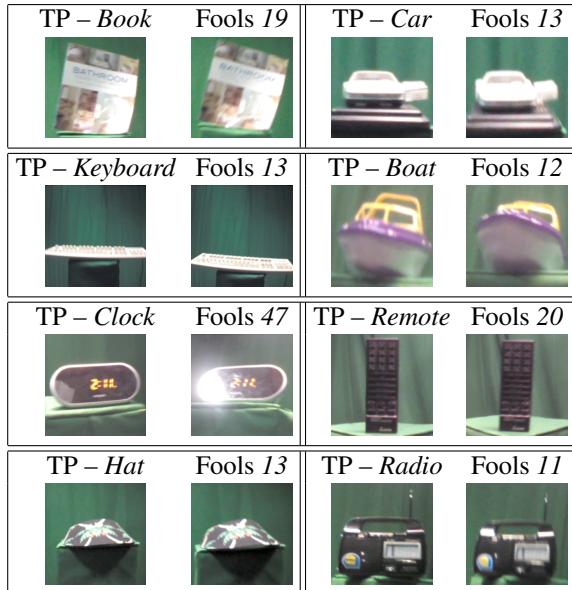


Table 7: Adversarial samples for PV SIP

