Adversarial Examples that Fool Detectors Supplementary Materials

Jiajun Lu*, Hussein Sibai*, Evan Fabry
University of Illinois at Urbana Champaign
{jlu23, sibai2, efabry2}@illinois.edu

The videos in the supplementary materials are in H.264 (MPEG-4 AVC) format, and it can be downloaded from http://www.divx.com/en/software/technologies/h264. We include partial data and comprehensive results for three experiment settings: attacking stop sign video sequences digitally, attacking physically printed stop signs and attacking face video sequences digitally.

1. Digital Experiments on Stop Signs

The materials related to attacking stop sign video sequences digitally can be found in folder "digital_ experiments_stop_sign". The "train" folder includes Faster RCNN's [1] (trained on MSCOCO [3]) response on the training data, as well as digitally attacked training data with three different perturbations.

- "train_original_for_stop1_stop2": Faster RCNN's response on all the training data for adversarial perturbation pattern 1 and pattern 2. Stop signs are reliably detected.
- "train_original_for_stop3": Faster RCNN's response on all the training data for adversarial perturbation pattern 3. Stop signs are reliably detected.
- "train_stop1_perturbation": Faster RCNN's response on all the training data with adversarial perturbation pattern 1 applied. Stop signs are not detected or detected as a wrong class.
- "train_stop2_perturbation": Faster RCNN's response on all the training data with adversarial perturbation pattern 2 applied. In most images, stop signs are not detected or detected as a wrong class.
- "train_stop3_perturbation": Faster RCNN's response on all the training data with adversarial perturbation pattern 3 applied. Stop signs are not detected or detected as a wrong class.

The "test_example" folder includes an original test example sequence, as well as these with three different adversarial perturbations applied digitally. We only visualize Faster RCNN's response on the "stop sign" class to keep the videos clean.

- "test_original_exp.mp4": Faster RCNN's response on one of the test videos. The stop sign is detected.
- "test_stop1_perturbation_exp.mp4": Faster RCNN's response on the same test video with adversarial perturbation pattern 1 applied. The stop sign is not detected.
- "test_stop2_perturbation_exp.mp4": Faster RCNN's response on the same test video with adversarial perturbation pattern 2 applied. The stop sign is occasionally detected.
- "test_stop3_perturbation_exp.mp4": Faster RCNN's response on the same test video with adversarial perturbation pattern 3 applied. The stop sign is not detected.

2. Physical Experiments on Stop Signs

The materials related to attacking physically printed stop signs can be found in folder "physical_experiments_stop_sign". The adversarial perturbations are generated from the digital experiments in the previous section. We first divide our results by backgrounds, because stop signs are detected differently with tree background and sky background. The "tree_background" folder includes Faster RCNN's response on six different physical adversarial stop signs with tree background: three different perturbations with two different brightness. In the videos, an adversarial stop sign (30 in x 30 in) is held by a person, and there is another real stop sign (20 in x 20 in) behind it. Some physical adversarial stop signs merge into background and are not detected, while some are detected.

- "not_detected/stop3_bright_tree_physical_exp.mp4": The bright adversarial perturbation pattern 3 with tree background is not detected by Faster RCNN.
- "occasionally_detected/stop2_dark_tree_physical_exp.mp4": The dark adversarial perturbation pattern 2 with tree background is occasionally detected by Faster RCNN.

*Both authors contributed equally

- "occasionally_detected/stop3_dark_tree_physical_exp.mp4": The dark adversarial perturbation pattern 3 with tree background is occasionally detected by Faster RCNN.
- "detected/stop1_bright_tree_physical_exp.mp4": The bright adversarial perturbation pattern 1 with tree background is detected by Faster RCNN.
- "detected/stop1_dark_tree_physical_exp.mp4": The dark adversarial perturbation pattern 1 with tree background is detected by Faster RCNN.
- "detected/stop2_bright_tree_physical_exp.mp4": The bright adversarial perturbation pattern 2 with tree background is detected by Faster RCNN.

The "sky_background" folder includes Faster RCNN's response on six different physical adversarial stop signs with sky background: three different perturbations with two different brightness. In the videos, an adversarial stop sign (30 in x 30 in) is held by a person, and there is another real stop sign (20 in x 20 in) parallel to it. It is much harder to create physical adversarial examples with sky background.

- "not_detected/stop3_bright_sky_physical_exp.mp4": The bright adversarial perturbation pattern 3 with sky background is not detected by Faster RCNN until the last second.
- "occasionally_detected/stop2_dark_sky_physical_exp.mp4": The dark adversarial perturbation pattern 2 with sky background is occasionally detected by Faster RCNN.
- "occasionally_detected/stop3_dark_sky_physical_exp.mp4": The dark adversarial perturbation pattern 3 with sky background is occasionally detected by Faster RCNN.
- "detected/stop1_bright_sky_physical_exp.mp4": The bright adversarial perturbation pattern 1 with sky background is detected by Faster RCNN.
- "detected/stop1_dark_sky_physical_exp.mp4": The dark adversarial perturbation pattern 1 with sky background is detected by Faster RCNN.
- "detected/stop2_bright_sky_physical_exp.mp4": The bright adversarial perturbation pattern 2 with sky background is detected by Faster RCNN.

3. Digital Experiments on Faces

The materials related to attacking face video sequences digitally can be found in folder "digital_experiments_face". The "train" folder includes Faster RCNN based face detector [2]'s response on the training data, as well as digitally attacked training data with three different perturbations.

- "train_original": Face detector's response on all the training data. Faces are reliably detected.
- "train_small_perturbation": Relatively small perturbations are applied to all the training data, and faces in the training data are not detected.
- "train_medium_perturbation": Medium perturbations are applied to all the training data, and faces in the training data are occasionally detected.
- "train_large_perturbation": Large perturbations are applied to all the training data, and faces in the training data are not detected.

The "test_example" folder includes an original test example sequence, as well as these with three different adversarial perturbations applied digitally.

- "test_original.mp4": Face detector's response on the original test video. The face is reliably detected.
- "test_small_perturbation.mp4": Face detector's response on the same test video with relatively small adversarial perturbations applied. The face is still detected in most frames (notice that this perturbation pattern fools face detector well digitally with the training images).
- "test_medium_perturbation.mp4": Face detector's response on the same test video with medium adversarial perturbations applied. The face is occasionally detected.
- "test_large_perturbation.mp4": Face detector's response on the same test video with large adversarial perturbations applied. The face is seldom detected.

References

[1] X. Chen and A. Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017. 1

[2] H. Jiang and E. Learned-Miller. Face detection with the faster r-cnn. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 650–657. IEEE, 2017. 2

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1