Adversarial Examples for Edge Detection: They Exist, and They Transfer Supplemental Material

Christian Cosgrove Alan L. Yuille
Department of Computer Science, The Johns Hopkins University
Baltimore, MD 21218 USA
ccosgro2@jhu.edu alan.l.yuille@gmail.com

1. Attacks shifting edges?

One potential concern is that these attacks are simply "shifting" the edges in an image away from the ground-truth edge. If this were the case, the attacks would be able to increase the loss while still roughly preserving the locations and shapes of edges in the image (because the in the ground-truth edges are only a few pixels wide, so a large loss can occur if the model produces edges only a few pixels away from the ground-truth edge). In a sense, this would mean the attacks were "cheating" on the attack problem.

Applying morphological filtering to the ground truth edge labels most likely reduces this problem (because it makes the ground-truth edges much thicker), but here we empirically verify that our attacks are not "cheating" in this way.

To check that our attacks are not simply shifting edges, we perform the attacks and measure the probability of detecting an edge as a function of distance from a true ground truth edge. We average these results over the BSDS500 test set. See the results in Figure 1.

If these attacks were simply shifting edges away from the ground truth edge, we would see a "bump" in the probabilities as distance increases. None of the attacks exhibit this behavior, so the attacks are not shifting edges.

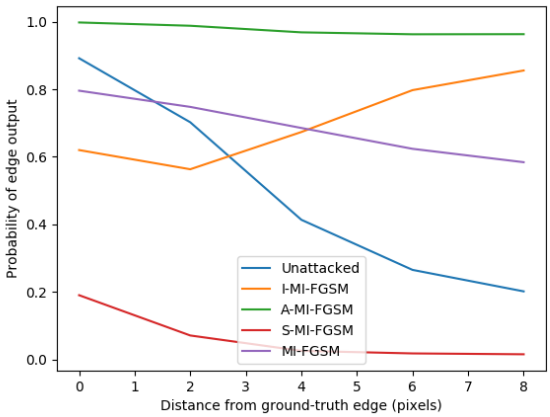
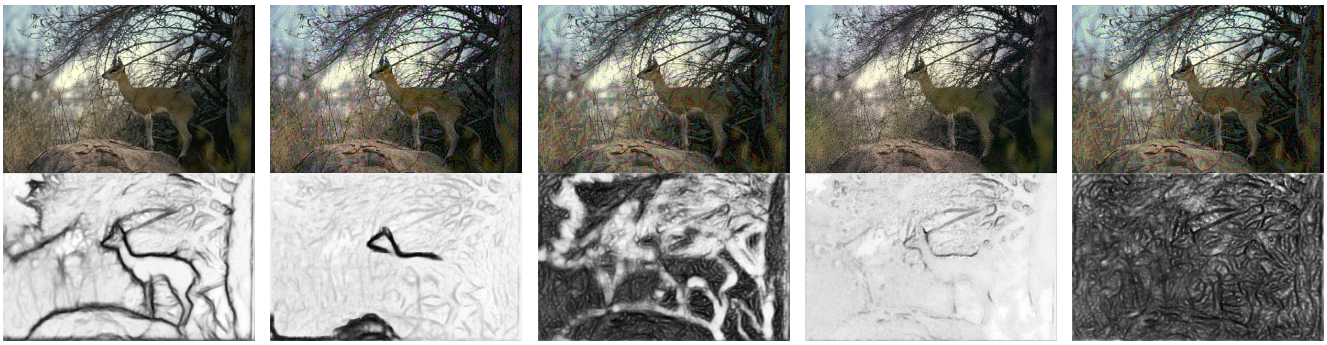


Figure 1



(a) no attack (b) MI-FGSM (c) I-MI-FGSM (d) S-MI-FGSM (e) A-MI-FGSM

Figure 2: More attacked images with $\epsilon = 16$.