Human Decisions on Targeted and Non-Targeted Adversarial Samples

Samuel M. Harding (hardinsm@indiana.edu)
Prashanth Rajivan (prajivan@andrew.cmu.edu)
Bennett I. Bertenthal (bbertent@indiana.edu)
Cleotilde Gonzalez (coty@cmu.edu)

Abstract

and efficient method for finding perturbations where, given a source image x, each of the 784 features representing the input is perturbed in the direction of the gradient by magnitude of e. e represents the magnitude of the perturbation. The strength of perturbation at every feature is limited by the same constant parameter e and the resultant is a adversarial stimuli $\tilde{x}$ of the original input x. With even small e it is possible to mislead such Deep Neural Networks (DNN) with a high success rate. Due to the nature of gradient descent on the loss function, it is not possible for the model to anticipate the outcome and therefore, the goal is to misclassify adversarial input $\tilde{x}$ as any other class than its correct class (y). Hence, it is a *non-targeted* form of attack.

Papernot et al (2016) proposed the Jacobian-based Saliency Map Attack (JSMA) to generate adversarial samples to mislead neural network model. This model used an iterative approach to modify a limited and specific set of features (among the 784 features) of the input image (x) for targeted misclassification. In this approach, an adversarial saliency map is calculated for the input image which contains the scores for each pixel that reflect how the pixel can help in achieving the intended target class ($\tilde{y}$) while reducing the probability of achieving any other class. Pixels with high saliency scores are perturbed by e repeatedly until the model misclassifies the input as the intended target class. Papernot et al. (2016) found that a deep neural network can be fooled with high success (97%) while only requiring small modifications (4.02%) of the input features of a sample; while humans identified 97.4% of the adversarial samples correctly and classified 95.3% of the adversarial samples correctly.

Adversarial Image Generation We quantified the amount of perturbation introduced by each algorithm by computing the L1-norm, or pixel-wise ($i; j$

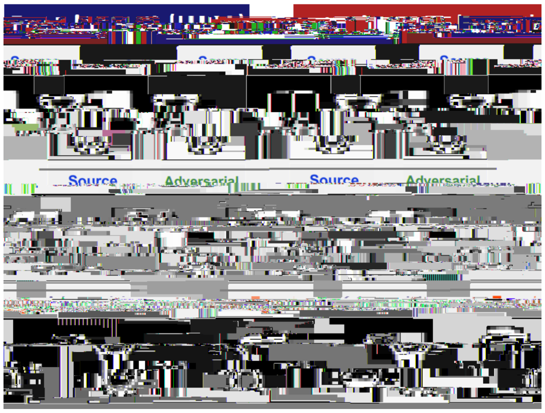

Figure 2: Examples of the image pairs shown in Experiments 1 (left columns) and Experiment 2 (right columns). In ‘Source-Source’ pairs, an MNIST digit was compared with itself; ‘Source-Adversarial’ pairs pitted an unperturbed MNIST digit against an adversarially-modified version of itself; finally ‘Target-Adversarial’ pairs compared an adversarial digit with an MNIST digit from the incorrect class produced by the DNN when classifying the adversarial image.

Experiment 1

In Experiment 1, we tested human classification, discrimination, and similarity judgments over images generated using the JSMA algorithm (targeted attack).

finding to a novel adversarial algorithm. The difference in accuracy when comparing across the two algorithms suggests that FGSM was more successful in confusing human judgments, perhaps due to the larger amount of perturbation, or the more global pattern of pixel changes.

Table 1: Classification Accuracy

	Experiment 1	Experiment 2
Unperturbed	96.8%	97.8%
Adversarial	94.2%	82.7%
Total	**95.5%**	**90.2%**

We next examined whether participants would correctly identify pairs of images showing the ‘same’ or ‘different’ digits, in spite of the adversarial modifications, in the Discrimination task. Overall accuracy was at 99.1% in Experiment 1, and 96.6% in Experiment 2 (see Table 2). A generalized, linear mixed-effects model over Trial Type (*Source-Source, Source-Adversarial, Target-Adversarial*) and Experiment (Experiment 1, Experiment 2) showed a significant main effects of Trial Type, $F(2,1794) = 71.937$, $p < .001$. There was also a main effect of Experiment, $F(1,1794) = 17.76$, $p < .001$, and a significant 2-way interaction, $F(2,1794) = 43.818$, $p < .001$. These results were driven primarily by better performance for the adversarial comparisons (*Source-Adversarial*, *Target-Adversarial*) in Experiment 1 than in Experiment 2, with no difference in *Source-Source* trials. This is consistent with the pattern of results found in the classification task, which showed that performance on images produced by the FGSM algorithm tended to be worse than over those generated by JSMA; furthermore, this is a novel demonstration that adversarial images can perturb human judgments in tasks other than Classification.

Table 2: Discrimination Accuracy

	Experiment 1	Experiment 2
Source-Source	99.9%	99.9%
Source-Adversarial	97.9%	95.0%
Target-Adversarial	99.7%	94.8%
Total	**99.1%**	**96.6%**

In similarity w33892

differences across the Experiments or image Type, using a linear mixed-effects model. Similarity ratings were significantly different across Trial Types; $F(2,1794) = 13{,}881$, $p < .001$. This difference was mostly in the *Source-Adversarial* and *Target-Adversarial* comparisons (see Figure 4). There was not a significant main effect of Experiment, $F(1,1794) = .712$, $p > .05$, but the interaction between Trial Type and Experiment was significant, $F(2,1794) = 46.627$, $p < .001$. This latter effect was due to the reversal in the two adversarial comparisons: while the ratings in *Target-Adversarial* pairs remained lower than the other comparisons, the additional noise introduced by FGSM seems to have made the adversarial image appear more similar to the intended target category than the procedure adopted by JSMA.

One possible explanation for this finding is that the distance between adversarial and source images was larger for stark differences in ratings as a function of trial type in the similarity task, we ran separate correlations for each stimulus type: *Source-Adversarial* similarity scores were significantly correlated with classification performance, $r(298) = .152$, $p < .01$, and marginally related to discrimination, $r(298) = .112$, $p = .053$. *Target-Adversarial* performance was likewise correlated between similarity, $r(298) = -.129$, $p < .05$, and discrimination, $r(298) = -.131$, $p < .05$. Finally, *Source-Source* similarity judgments were only related to discrimination performance, $r(298) = .272$, $p < .001$.

In Experiment 2, individual performance in the classification and discrimination tasks was significantly correlated, $r(298) = 0.839$, $p < .001$. Separate correlations by stimulus type in the similarity task showed that *Target-Adversarial* judgments were significantly negatively correlated with classification performance, $r(298) = -.328$, $p < .001$, and related to discrimination, $r(298) = -.471$, $p < .001$. *Source-Adversarial* performance was correlated between similarity, $r(298)$ and discrimination, $r(298) = .129$, $p < .05$.

Together, these results suggest that the different tasks rely on similar perceptual representations, and that individuals' performance on one task could be used to predict their abilities in the other domains. If, for example, a subject rates adversarial images as particularly dissimiliar to their unperturbed counterparts, they may be less prone to incorrectly classify the image, and therefore be less vulnerable to these types of perturbations, making the collection of explicit similarity ratings an important tool for assessing the risk posed by adversarial images.

Figure 4: Mean similarity ratings across Experiments 1 (blue) and 2 (red), separated by the image pair shown to subjects.

General Discussion

Current research on AML claims that humans are insensitive to the perturbations introduced in adversarial samples; however, these claims are not based on evidence from empirical research. This study represents the systematic attempt to test humans susceptibility to adversarial stimuli, and the results suggest that previous claims may have been overstated. Although adversarial stimuli are very effective in fooling on AMLhadvT(261 6bSls(AMLha1.955 2615 Td [(cl(AMLharforma relations were generally very small accounting for no more