Exploring and Expanding the One-Pixel Attack
Umair Khan, Walt Woods, Christof Teuscher

Abstract: In machine learning research, adversarial examples are normal inputs to a classifier that have been specifically perturbed to cause the model to misclassify the input. These perturbations rarely affect the human readability of an input, even though the model’s output is drastically different. Recent work has demonstrated that image-classifying deep neural networks (DNNs) can be reliably fooled with the modification of a single pixel in the input image, without knowledge of a DNN’s internal parameters. This “one-pixel attack” utilizes an iterative evolutionary optimizer known as differential evolution (DE) to find the most effective pixel to perturb, via the evaluation of numerous candidate solutions with a specific fitness function. We first improve upon the original implementation of the attack by designing a fitness function to minimize the magnitude of the perturbation in addition to the network confidence. The original attack achieves a success rate of 37% on our basic model with a mean attack RMSE of 0.02418; the improved attack achieves a success rate of 38% with a mean attack RMSE of 0.01946. We then explore the attack’s efficacy by comparing its performance in neural networks of different depths, and analyze the technique by computing per-pixel heatmaps of vulnerabilities in input images. Our findings highlight the applicability of the technique across networks, while at the same time demonstrating the shortcomings of DE in maximizing the attack potential (Figure 1). Future work could address these shortcomings, as well as extend the one-pixel attack to new domains (e.g. video).

Figure 1 -- A heatmap analysis of DE. The strength of a region indicates the vulnerability of the image in that area. Note that in the final case, DE fails to find the most effective pixel to perturb.