Fooling a CNN to Misclassify Images

James Wu

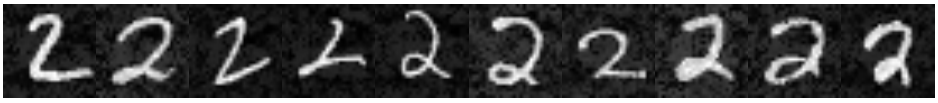
March 17^{th} 2017

Image Visualization

Figure 1: Original Images

Figure 2: Noise Images

Figure 3: Adversarial Images



Comments

Although the adversarial images are slightly different from the original images, it is clear that they are still images of two and not six. This demonstrates one of the limitations of convolutional neural networks for classification.

During the training of the CNN and the process of generating the adversarial images, it was noted that when the neural network was trained for more iterations, more iterations would also be required when generating adversarial images increased as well. This indicates that it was "more difficult" to generate an adversarial image when the CNN was trained for a longer period of time. This can be indicative of a possible solution to the problem, where a large data set would be required to properly train a CNN such that they cannot be easily "fooled" to misclassify.