Jorge Martinez

Vincent Sylvester

March 5, 2020

Final Project Proposal

Based on the findings of Szegedy et al. [1], we will attempt to create adversarial examples to iteratively improve the accuracy of a classifier, created by ourselves, on the MNIST dataset (and potentially others). From here, we will compare our classifier to a state-of-the-art network and the current record holder on Kaggle.

Adversarial examples are, in the case of image recognition, visually imperceptible perturbations that cause images to be misclassified. There are several ways that we can find these images, but mostly it is trying to find the tiniest change to the input image that results in a misclassification. Since finding the minimum perturbation is a hard problem, we will test a few approximation methods for finding these images, including L-BFGS, which is suggested by the authors [1]. Then, we will create a pool of adversarial examples that can then be mixed into our training data to try and improve accuracy. The question we hope to answer is how many times, on average, you can iterate this process (perturbing an image and then using it for training) before it ceases to have any learning value. Essentially, at some point, the perturbation must become large enough that the image is no longer close enough to the actual picture to be useful.

The usefulness of this approach is that it may decrease our reliance on the data, robust our network, and make it more durable against adversarial perturbations. Our first step will be to find a pre-trained classifier and try to implement the algorithm to create adversarial examples.

References

1. Szegedy, C., Zaremba, W., Sutskever, I., Bruna Estrach, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks*. Paper presented at 2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada.