Project Statement

IFT6390

Gabriel C-Parent, Dora Fugère

Following the recent interest in messing with classifier systems i.e. the generation of adversarial examples in input space (Szegedy et al. 2014, Goodfellow et al. (2014)), we got interested in applying the same treatment to simpler classifiers.

We targetted the simplest classifier we knew of, the linear regression with stochastic gradient descent (one-hot encoding and argmax to decide the class). We used the MNIST dataset to perform all experiments. This was motivated by its popularity in machine learning litterature (http://yann.lecun.com/exdb/mnist/) and because images are fun to mess with.

The objective of this research was twofold, first we wanted to observe the distribution of minimal distortion necessary for misclassification of correctly classified inputs (the minimal distance to an adversarial example). We were really intrigued to compare it to the values obtained in (Szegedy et al. 2014) for neural networks.

Our second objective was to investigate the interplay of the regularization schemes (or lack thereof) on the ease of creation of adversarial examples.

Bibliography

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Nets." *Technical Report ArXiv:1406.2661.*

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. 2014. "Intriguing Properties of Neural Networks." *ICLR.*