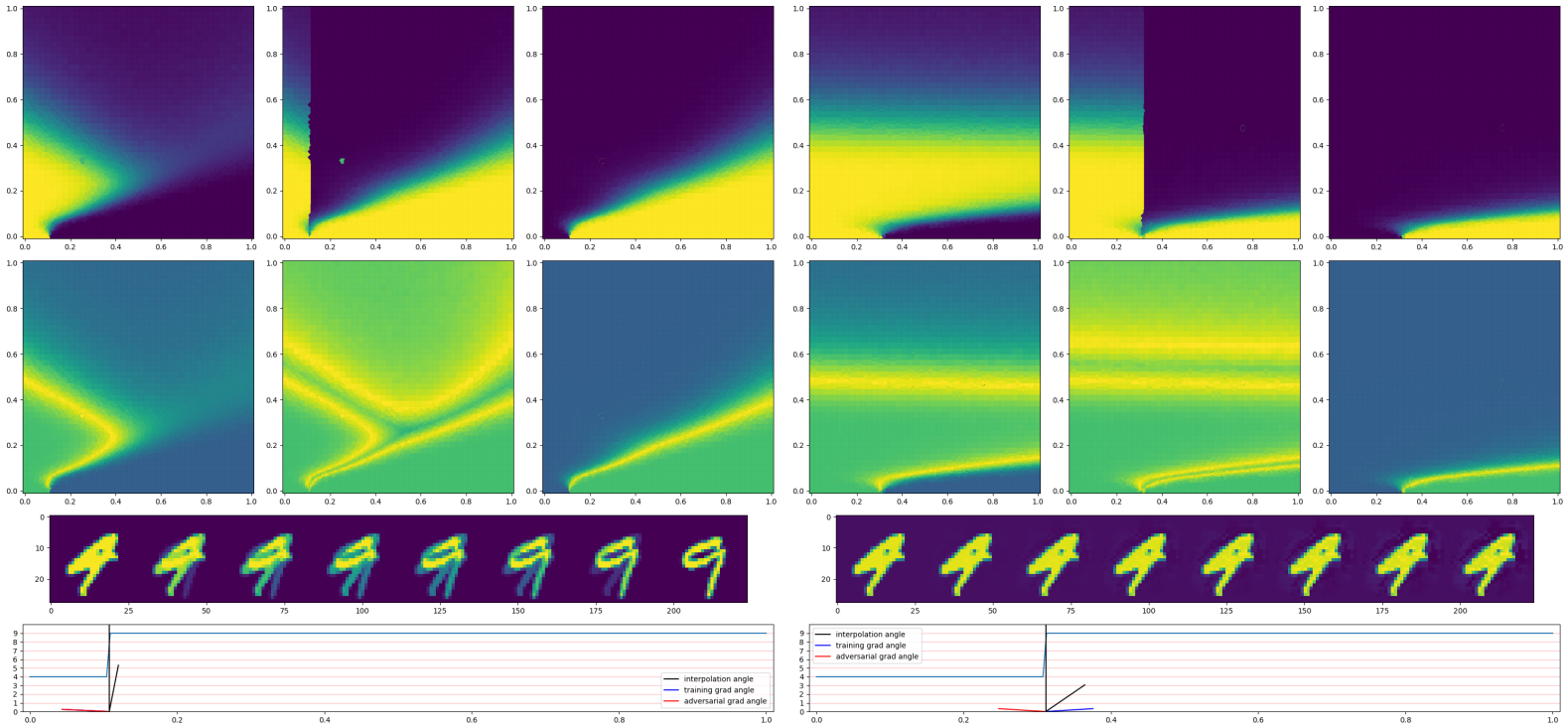
A Geometric Framework for Adversarial Vulnerability in Machine Learning



Brian Bell : University of Arizona

November 6, 2023

Contents

Intriguing Properties of Neural Networks Szegedy et al. (2014)

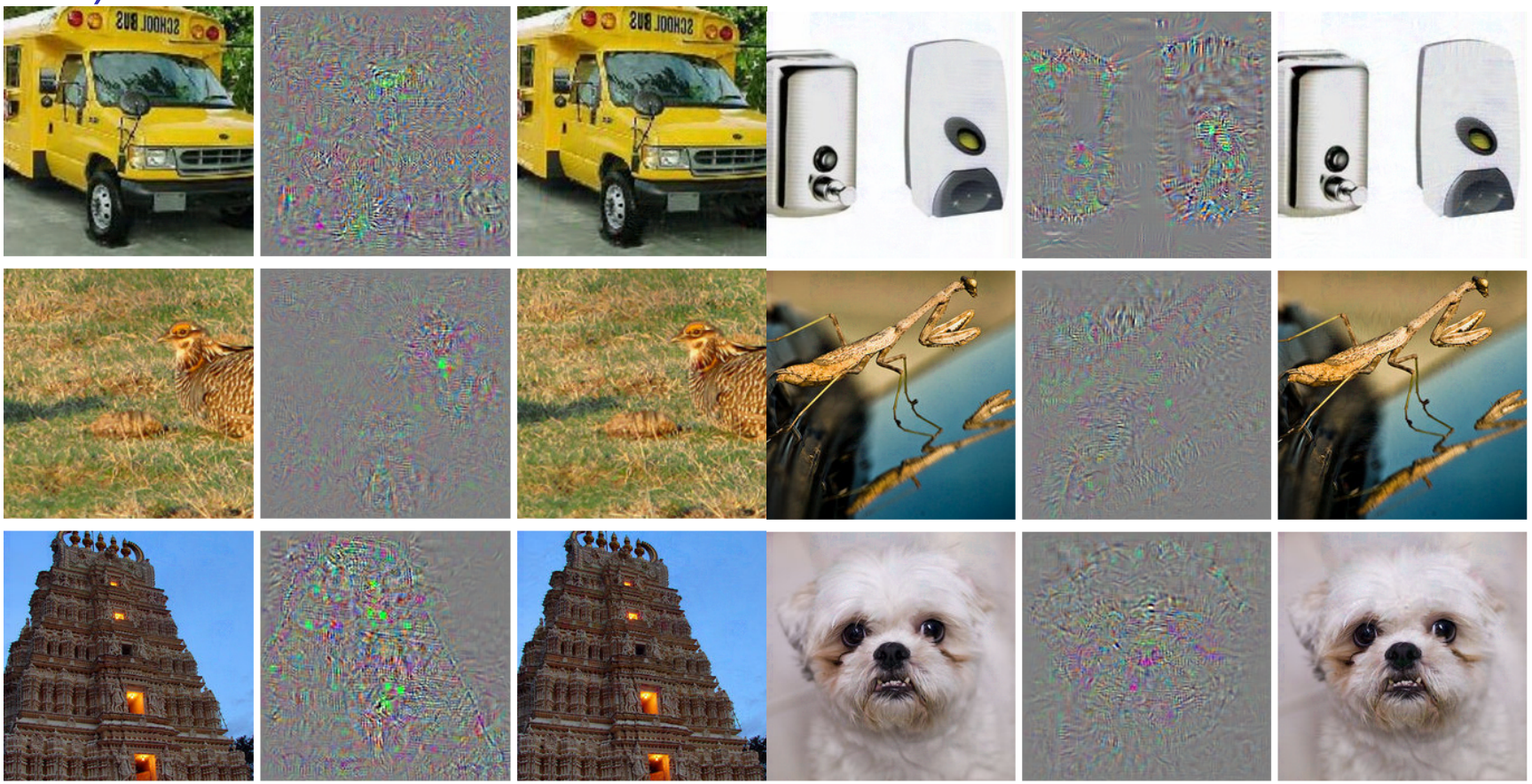


Figure: Natural Images are in columns 1 and 4, Adversarial images are in columns 3 and 6, and the difference between them (magnified by a factor of 10) is in columns 2 and 5. All images in columns 3 and 6 are classified by AlexNet as "Ostrich" Szegedy et al. (2014)

Attacks : L-BFGS

Let $f : \mathbb{R}^m \to \{1, ..., k\}$ be a classifier and assume f has an associated continuous loss function denoted by $\text{loss}_f : \mathbb{R}^m \times \{1, ..., k\} \to \mathbb{R}^+$ and l a target adversarial .

Minimize $\|r\|_2$ subject to:

1. $f(x + r) = l$
2. $x + r \in [0, 1]^m$

The solution is approximated with L-BFGS as implemented in Pytorch or Keras. This technique yields examples that are close to their original counterparts in the L^2 sense.

Attacks : L-BFGS : MNIST

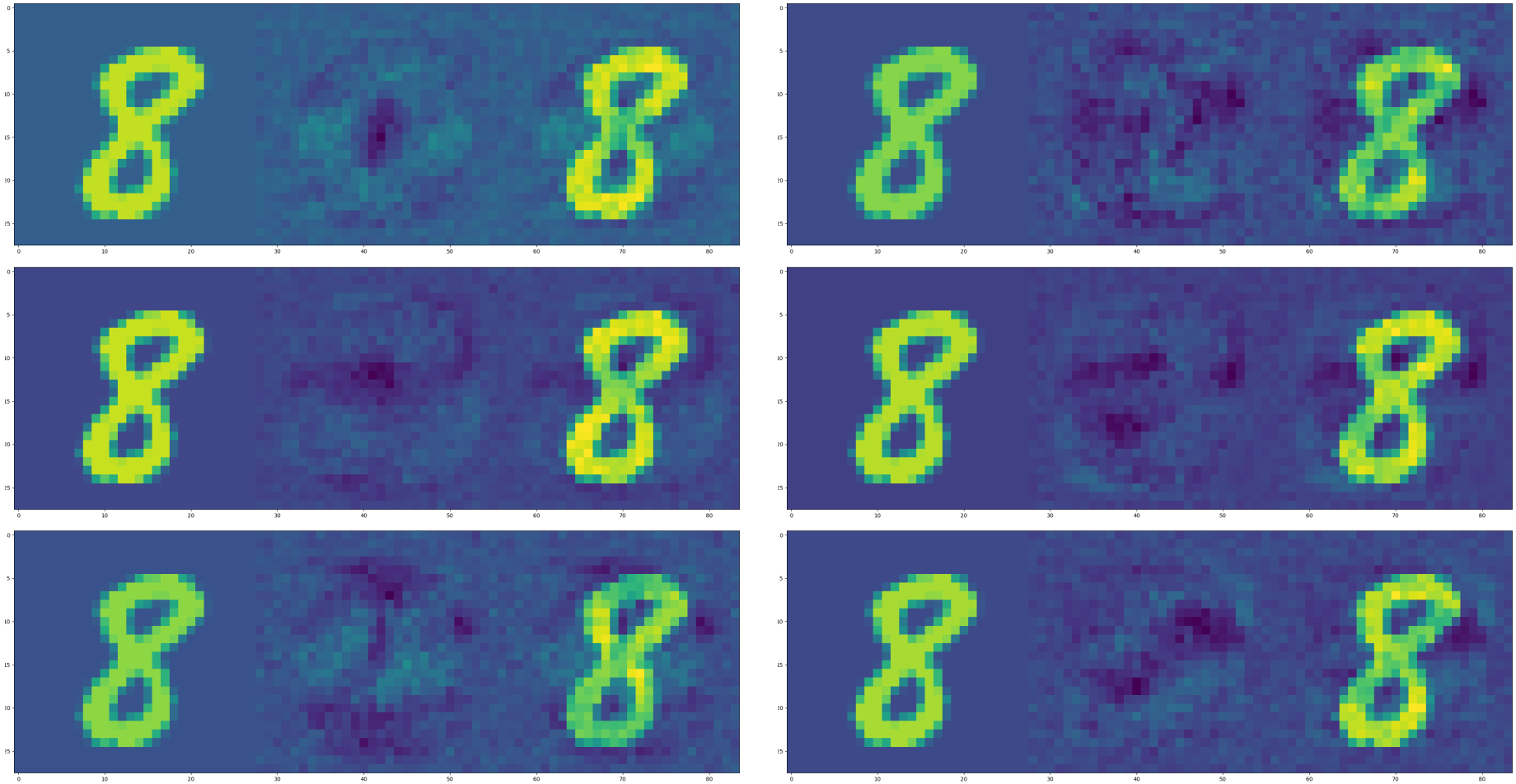


Figure: Original images on the left, Perturbation is in the middle, Adversarial Image (total of Original with Perturbation) is on the right. Column 1 shows an original 8 being perturbed to adversarial classes 0, 2, and 4. Column 2 shows adversarial classes 1, 3, and 5

Attacks : Distortion

Borrowing a metric from Szegedy et al to compare the magnitude of these distortions, we will define

Definition

Distortion is the L^2 norm of the difference between an original image and a perturbed image, divided by the square root of the number of pixels in the image:

$$\sqrt{\frac{\sum_i \hat{(}x_i - x_i)^2}{n}}$$

Distortion is L^2 magnitude normalized by the square-root of the number of dimensions so that values can be compared for modeling problems with differing numbers of dimensions.

Attacks : L-BFGS : MNIST

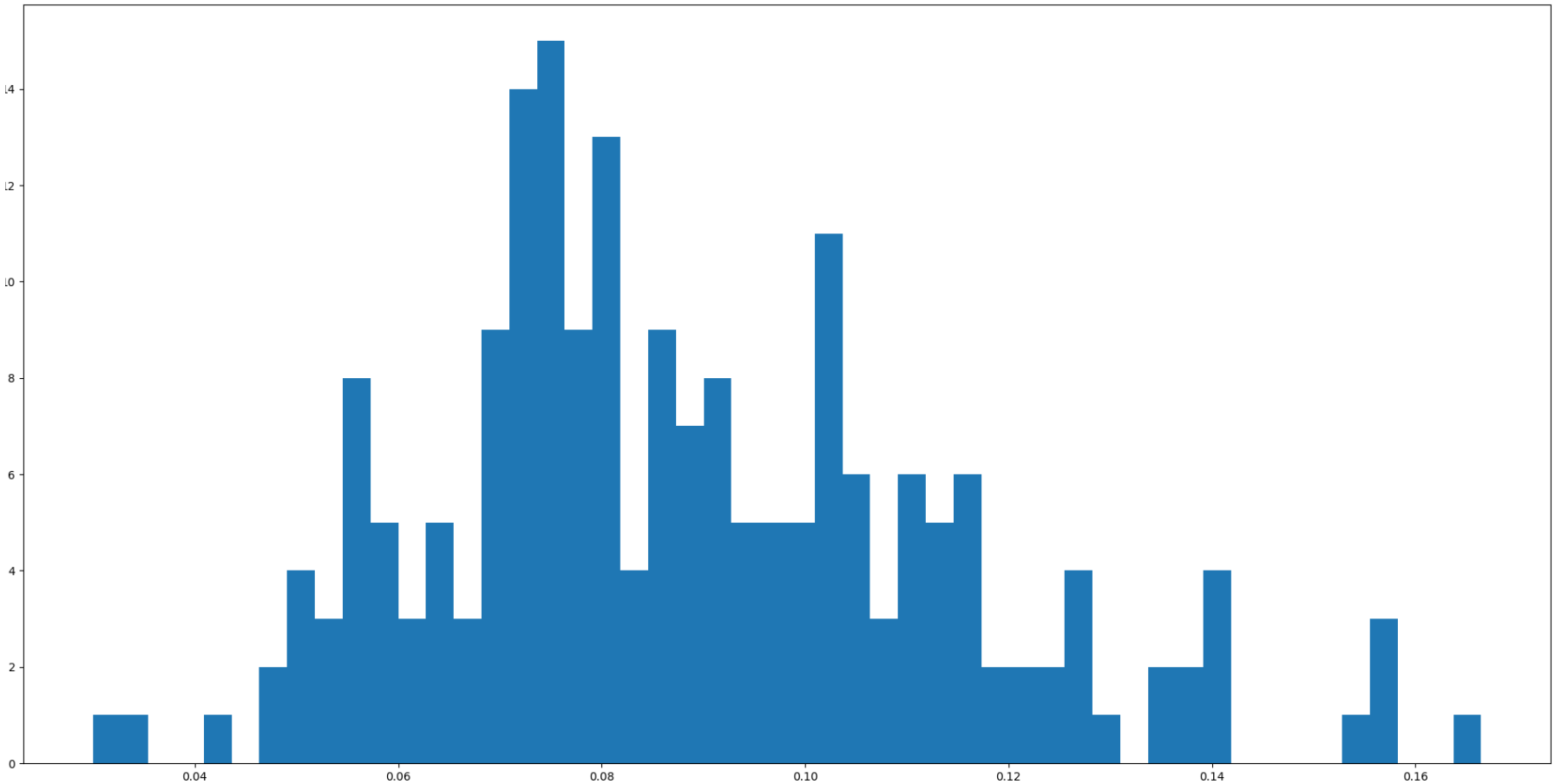


Figure: A histogram of the distortion measured for each of 900 adversarial examples generated using L-BFGS against the FC-200-200-10 network on Mnist. Mean distortion is 0.089.

Attacks : L-BFGS : ImageNet

Figure: Original images on the left, Perturbation (magnified by a factor of 100) by is in the middle, Adversarial Image (total of Original with Perturbation) is on the right. Adversarial classes are Burrito, Bison, Taxi, and Paddle Wheel (Top Left, Top Right, Bottom Left, Bottom Right)

Attacks : L-BFGS : ImageNet

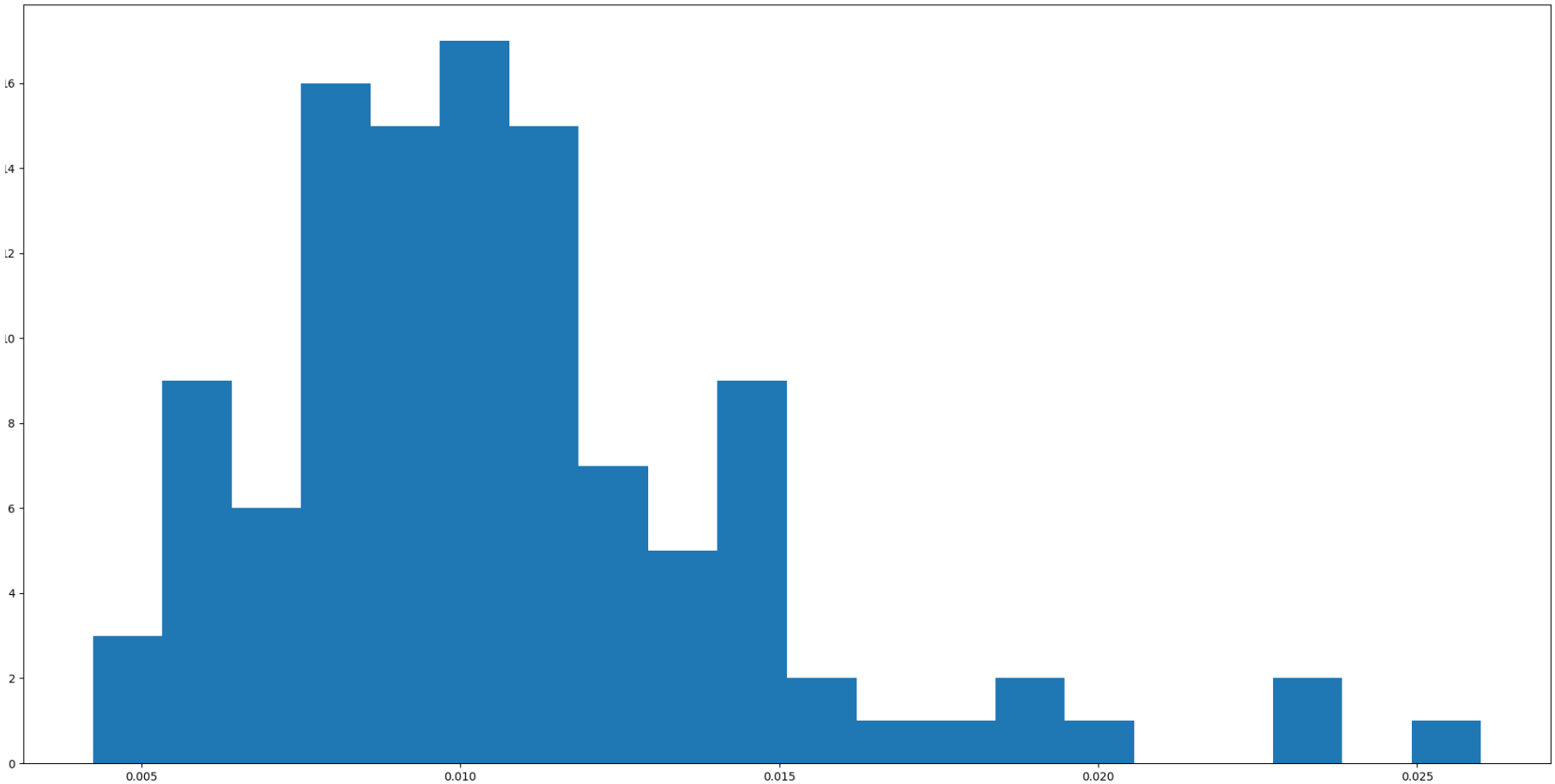


Figure: A histogram of the distortion measured for each of 112 adversarial examples generated using L-BFGS against the VGG16 network on ImageNet images with mean distortion 0.0107

Attacks : FGSM

A single step attack process using the gradient of the loss function L with respect to an image to find the adversarial perturbation Goodfellow et al. (2014). for given ε, the modified image $\hat{x}$ is computed as

$$\hat{x} = x + \epsilon \text{sign}(\nabla L(P_w(x), x)) \tag{1}$$

This method is simpler and much faster to compute than the L-BFGS technique described above, but produces adversarial examples less reliably and with generally larger distortion.

Attacks : IGSM

In Kurakin et al. (2016) an iterative application of FGSM was proposed. After each iteration, the image is clipped to a εL_∞ neighborhood of the original. Let $x'_0 = x$, then after m iterations, the adversarial image obtained is:

$$x'_{m+1} = \text{Clip}_{x,\epsilon}\Big\{ x'_m + \alpha \times \text{sign}(\nabla \ell(F(x'_m), x'_m)) \Big\} \tag{2}$$

This method is faster than L-BFGS and more reliable than FGSM but still produces examples with greater distortion than L-BFGS.

Attacks : IGSM : ImageNet

Figure: adversarial example generated against VGG16 (ImageNet) with IGSM. Original Image (Rose Hip) on the left, adversarial image (Baseball) on the right.

Defining Adversarial Examples

Definition

Consider a point $x \in X$ with corresponding class $c \in C$ and a classifier $\mathcal{C} : X \to C$. We say that x admits an $(\varepsilon, d)-$*adversarial example* on $\mathcal{C}$ if there exists a point $\hat{x}$ such that $d(x, \hat{x}) < \varepsilon$ and $\mathcal{C}(\hat{x}) \neq c$.

This definition refers to the most general case of intentional mis-classification. The adversarial class can also be explicitly targeted:

Definition

Consider a point $x \in X$ with corresponding class $c \in C$ and a classifier $\mathcal{C} : X \to C$. We say that x admits an $(\varepsilon, d, c_t)-$*targeted adversarial example* on $\mathcal{C}$ if there exists a point $\hat{x}$ such that $d(x, \hat{x}) < \varepsilon$ and $\mathcal{C}(\hat{x}) = c_t$.

Citations I

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *arXiv:1412.6572 [cs, stat]*. arXiv: 1412.6572.

Kurakin, A., Goodfellow, I., and Bengio, S. (2016). Adversarial examples in the physical world. *arXiv:1607.02533 [cs, stat]*. arXiv: 1607.02533.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014). Intriguing properties of neural networks. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.