Adversarial Examples Are a Natural Consequence of Test Error in Noise

1. Introduction

The machine learning community has researchers working on each of these two types of errors: adversarial example researchers seek to measure and improve robustness to small-worst case perturbations of the input while corruption robustness researchers seek to measure and improve model robustness to distributional shift.

Given that successful adversarial defense methods should also improve some types of corruption robustness we recommend that future researchers consider evaluating corruption robustness in addition to adversarial robustness.

2. Related Work

3. Adversarial and Corruption Robustness

Both adversarial robustness and corruption robustness can be thought of as functions of the error set of a statistical classifier.

4. Errors in Gaussian Noise Suggest Adversarial Examples

We will start by examining the relationship between adversarial and corruption robustness in the case where q consists of images with additive Gaussian noise.。

**The Linear Case.** For linear models, the error rate in Gaussian noise exactly determines the distance to the decision boundary.

**Comparing Neural Networks to the Linear Case.**

**Visual Confirmation of the Half-space Model**

5. Concentration of Measure for Noisy Images

concentration of measure

The Gaussian Isoperimetric Inequality

6. Evaluating Corruption Robustness

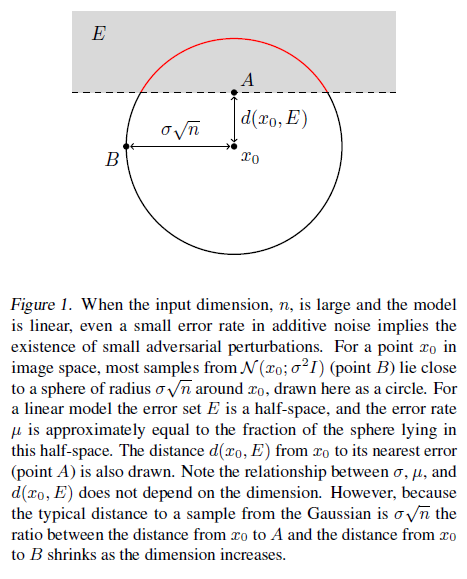
7. Conclusion

This paper investigates whether we should be surprised to find adversarial examples as close as we do, given the error rates we observe in corrupted image distributions. After running several experiments, we argue that the answer to this question is no.

1. The nearby errors we can find show up at the same distance scales we would expect from a linear model with the same corruption robustness.

2. Concentration of measure shows that a non-zero error rate in Gaussian noise logically implies the existence of small adversarial perturbations of noisy images.

3. Finally, training procedures designed to improve adversarial robustness also improve many types of corruption robustness, and training on Gaussian noise moderately improves adversarial robustness.



**Comparing Neural Networks to the Linear Case**

**Visual Confirmation of the Half-space Model**

5. Concentration of Measure for Noisy Images