Adversarially Robust Generalization Requires More Data

Introduction:

This paper shows the difference between standard generalization and robust generalization and proves that the sample complexities can be considerably different from one another. This describes and proves that the accuracy of the training depends on the size of the data set that is provided to the model. With adversarially training there needs to be more data than of standard training to have the same outcome. Which brings up the following question that the researchers propose:

> *"How does the sample complexity of standard generalization compare to that of adversarially robust generalization?"*

They answer this question by looking at the distributions of Bernoulli and Gaussian and by analyzing the robust generalization for them both. This is looked at by subsampling the dataset at various rater and study the impact of each sample size on their adversarial robustness. For this review, we will only be looking at the MNIST dataset rather than MNIST, SVHN and CIFAR10 all together, and the goal is to learn a classifier that achieves good test accuracy even under l-infinity-bounded perturbations.

Going further, most of their analysis was to create a lower bound for the two distributions in order to show the number of samples required for robust generalization. This lower bound is important and provides a hardness that shows that any similar distributions will follow the same hardness. A good standard error can be achieved from a single sample whereas the robust generalization approach requires a significant more samples to provide the same error which is difficult to find the right number of samples for each dataset. Thus, no algorithm can produce a robust classifier without many samples.

Main Result:

We will begin by defining some important terms that help with understanding the topic of this paper further, such as adversarial examples, robust learning and specifically what a perturbation is. Which are defined as the following:

- *Adversarial examples:* machine learning models misclassify examples that are only slightly different from correctly classified examples drawn from the data distribution
- *Robust learning:* is learning that either or both better sense making through deep conceptual understanding and fast and accurate procedural fluency
- *Perturbation:* method for solving a problem by comparing it with a similar one for which the solution is known

https://arxiv.org/pdf/1804.11285.pdf

https://en.wikipedia.org/wiki/Adversarial_machine_learning

https://learnlab.org/research/wiki/Robust_learning

https://www2.isye.gatech.edu/~nemirovs/FullBookDec11.pdf

- This book is devoted to Robust Optimization — a specific and relatively novel methodology for handling optimization problems with uncertain data
- Chapter 1
-