Towards Deep Learning Models Resistant to Adversarial Attacks

Aleksander Madry, Aleksandar, Markelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu

Presented by: Ethan Rathbun

Problem Statement

- Adversarial examples propose a fundamental threat to ML models
- Improving and further understanding the limitations of model robustness is important
- Defenses against a generalized adversary are most useful

The Saddle Point Problem

- They formulate a “minimax” problem over the set of possible model parameters and attack perturbations.

- The goal is to find a model that minimizes the expected loss under adversarial perturbations bounded within an l_p ball.

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}} L(\theta, x+\delta, y) \right]$$

- This is the same framework commonly used in zero-sum games

Approximating the Inner Maximization

- Given a fixed model we can approximate a local solution to the saddle point problem by using a sufficiently strong adversarial attack.

- In this paper they essentially propose using vanilla PGD as the foundation of their optimization.

$$x^{t+1} = \Pi_{x+\mathcal{S}} \left(x^t + \alpha \operatorname{sgn}(\nabla_x L(\theta, x, y)) \right)$$

Adversarial Training

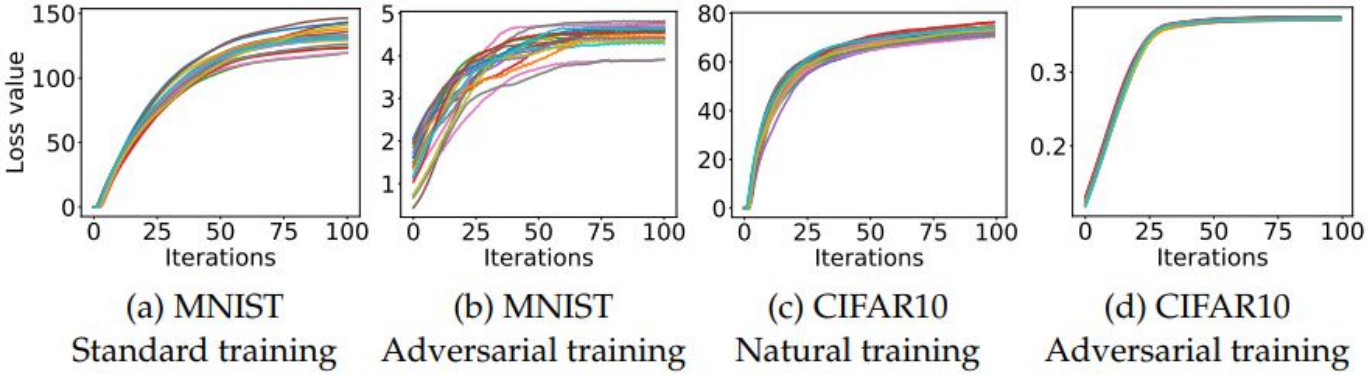
- The core concept of adversarial training is to train a model against adversarial samples in order to improve robustness.

- From my reading it seems they replace the training set with adversarial examples once and then train on them, but this is unclear.

Towards Universally Robust Networks

- The paper concedes that both the inner and outer maximization problems are highly non-convex, and thus exact solutions are intractable.

- In spite of this they show empirically that PGD can find good approximations to the inner maximization problem.

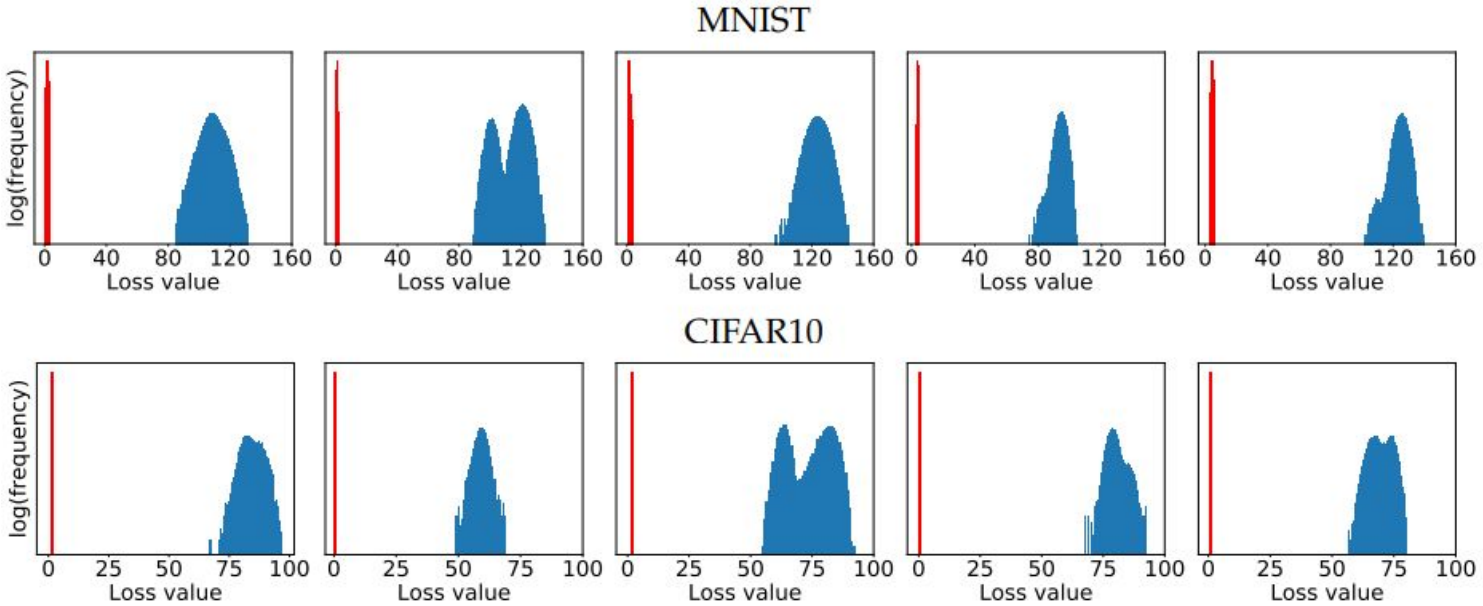


(a) MNIST Standard training | (b) MNIST Adversarial training | (c) CIFAR10 Natural training | (d) CIFAR10 Adversarial training

Here they run PGD multiple times with different starting points on the same natural example.

More Experiments

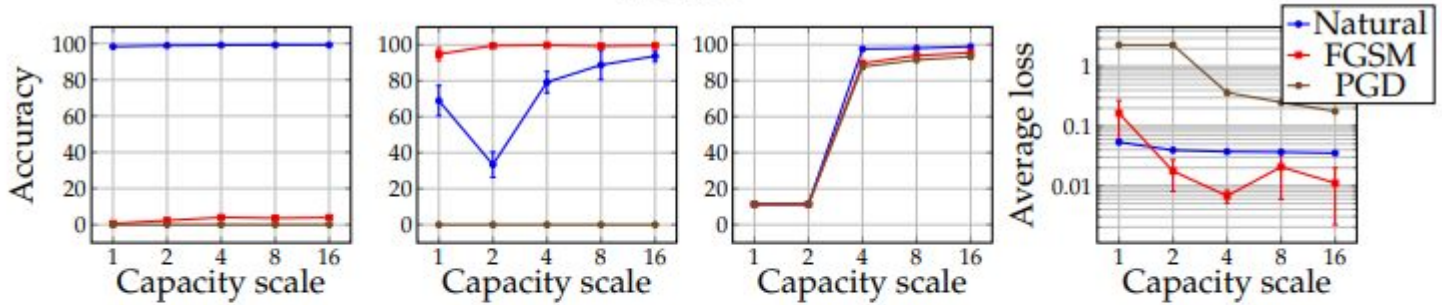
Here they attack vanilla (blue) and adversarially trained (red) models with PGD on 5 different samples. They run the experiment 10^5 times and plot the histograms of the loss.



Model Complexity

They generally show that more complex models perform better in the outer minimization problem

MNIST



CIFAR10

	Simple	Wide
Natural	92.7%	95.2%
FGSM	27.5%	32.7%
PGD	0.8%	3.5%

(a) Standard training

	Simple	Wide
Natural	87.4%	90.3%
FGSM	90.9%	95.1%
PGD	0.0%	0.0%

(b) FGSM training

	Simple	Wide
Natural	79.4%	87.3%
FGSM	51.7%	56.1%
PGD	43.7%	45.8%

(c) PGD training

	Simple	Wide
Natural	0.00357	0.00371
FGSM	0.0115	0.00557
PGD	1.11	0.0218

(d) Training Loss

Limitations

- Writing
- Experiments
- Method

Strengths

- Writing

- Experiments

- Methods

Discussion and Questions