Towards Deep Learning Models Resistant to Adversarial Attacks

01/31/2020

Presenter: Zijie Pan

https://qdata.github.io/deep2Read/

Motivation

Although trained networks are good on classifications on benign examples, the adversary is often able to manipulate the input so that the model produces an incorrect output.

Such phonenomen draws attentions to deep neural network design and training.

Many attack and defense mechanisms having been proposed:
defensive distillation, feature squeezing etc.

Problem: can not guarantee that find the most adversarial example, hard to find the extent of adversarial attacking that the model can resist

Thus: To put the attack and defense into a framework.

Claim / Target Task

- Adversarial attacking for neural network is essentially a min-max problem

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

- Model capacity plays an important role

Contribution 1:

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

1. Naive classification is normally construed as *empirical risk minimization*: just minimizing the expected loss for input-output pairs drawn from a data distribution *D*, which ceases to perform well once inputs are chosen adversarially.

1. Meaning of the formula: find the parameters θ that minimize the expected adversarially perturbed loss (which itself is found by maximizing the loss respect to the perturbation δ).

Contribution 1 background:

Choose of Inner optimizer:

fast gradient sign method (FGSM) represents a single step in the inner optimization

iterated projected gradient descent (PGD), represents a stronger inner optimizer

Back to the framework:

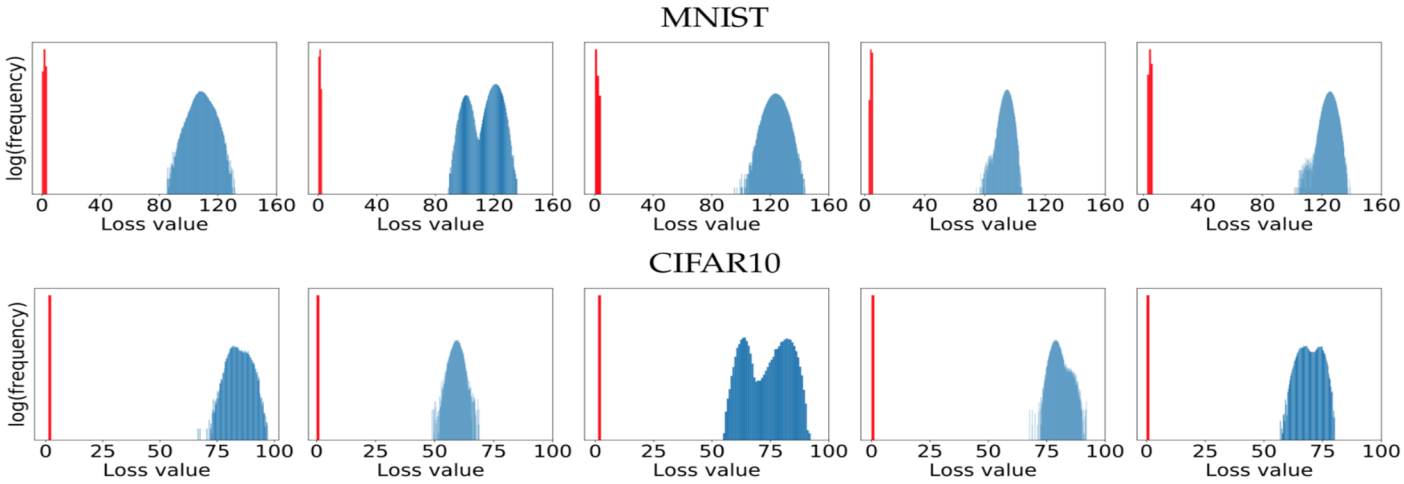
Saddle point problem,involves tackling both a non-convex outer minimization problem and a non-concave inner maximization problem

Contribution 1 Cont.

Saddle point problem solving:

Inner Part:

- Using *MNIST* and *CIFAR10* datasets
- the authors present empirical evidence that, while global maxima cannot in general be found, local maxima are all of similar quality.
- Starting PGD from 10^5 random points yielded loss with a tightly concentrated distribution with no observed outliers and all pairs of them are relatively independent



Contribution 1 Cont.

With valid target function, whether it's reasonable to use gradient information from the inner (maximization) problem as a valid update in the outer (minimization) problem.

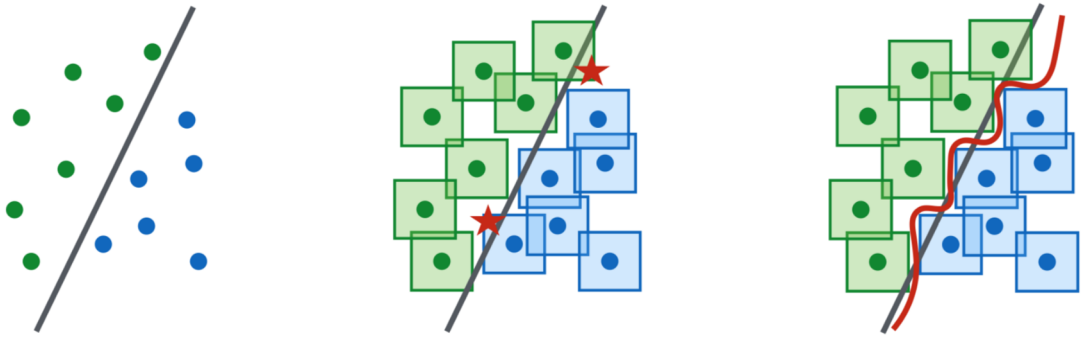
Danskin's theorem states this is indeed true and gradients at inner maximizers corresponds to descent directions for the saddle point problem

Since problem is not differentiable, the theorem assumption does not satisfy. However, empirically it works.

Conclusion 1: Using the framework above can train DNN that is robust (resistant to adversarial attacking)

Conclusion 2:

Model capacity effects robustness: classifiers must be significantly higher-capacity to be robust to adversarial examples



Conclusion 2 Cont.

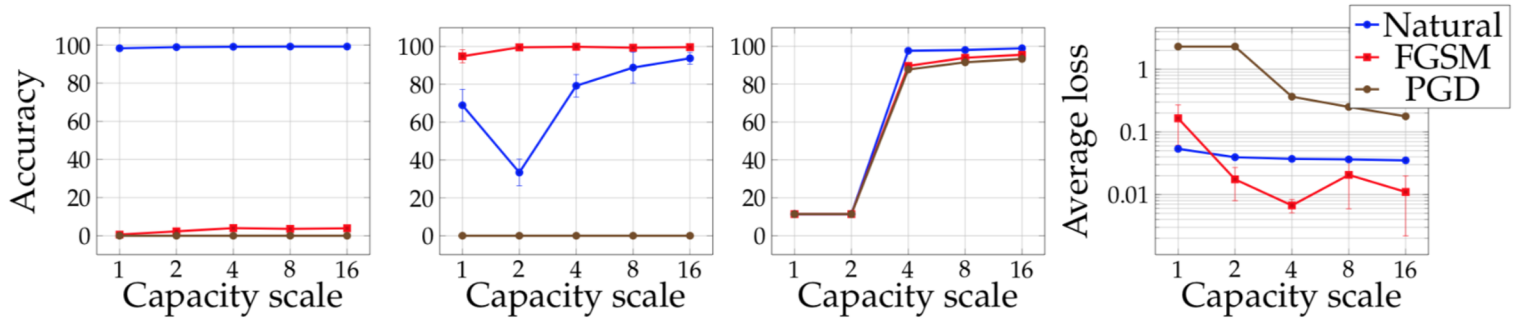
Conclusion draw from settings of varying capacity and different levels of adversarial training (none, FSGM only, and PGD):

- Capacity alone helps
- FSGM training only helps against FSGM-generated adversarial examples, but not against stronger attacks (Overfitting)
- Small-capacity models can't classify natural images at all when trained with PGD adversaries (large-capacity models can)
- The value of the saddle point problem decreases as we increase the capacity
- More capacity and stronger adversaries decrease transferability.

Experiment Result

Network Capacity on network performance:

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

Experiment Result Cont.

MNIST:

ε = 0.3,two convolutional layers with 32 and 64 filters respectively, each followed by 2 × 2 max-pooling, and a fully connected layer of size 1024.

Results on natural examples are good, but not on FGSM and PGD adversarial examples

Investigation in Appendix C: first convolutional layer only had 3 active filters and softmax output biases were skewed

CIFAR10:

- Original ResNet and its 10× wider variant
- Adversarial robustness of our network is significant but still far from satisfactory

Conclusion and Future Work

- Min-max framework is a good start

- Model capacity affects the model performance and it's a direction for future work

References