Convergence of Adversarial Training in Overparametrized Neural Networks

Ruiqi Gao[*,1], Tianle Cai[*,1], Haochuan Li[2],
Liwei Wang[1], Cho-Jui Hsieh[3], Jason D. Lee[4]

[1] Peking University, [2] MIT, [3] UCLA, [4] Princeton University
[*] Joint first author.

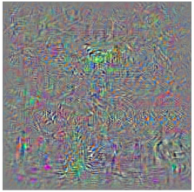
NeurIPS 2019

Introduction

Deep learning models are vulnerable to *adversarial attacks*.

(a) Schoolbus

$+0.1\times$



(b) Perturbation

$=$

(c) Ostrich

Figure: Szegedy et al. (2014)

Introduction(cont.)

- Most common white-box defenses are based on *adversarial training*, that is, at each step we perform gradient descent on the loss evaluated at the adversarially perturbed data.

Introduction(cont.)

- Most common white-box defenses are based on *adversarial training*, that is, at each step we perform gradient descent on the loss evaluated at the adversarially perturbed data.
- We give the first proof of convergence of adversarial training based on sufficiently wide networks.

Introduction(cont.)

- Most common white-box defenses are based on *adversarial training*, that is, at each step we perform gradient descent on the loss evaluated at the adversarially perturbed data.
- We give the first proof of convergence of adversarial training based on sufficiently wide networks.
- Our analysis leverages recent work on Neural Tangent Kernel (NTK), combined with motivation from online-learning, and the expressiveness of the NTK kernel in the ℓ_∞-norm.

Setting

Formalizing the problem:

- Neural network $f(W, x)$.

Setting

Formalizing the problem:

- Neural network $f(W, x)$.
- Adversarial attack (PGD, FGSM, etc.) $\mathcal{A}(W, x) = x' \in \mathcal{B}(x)$
  ($\mathcal{B}(x)$ is the allowed perturbation set e.g. ℓ_2 or ℓ_∞ ball centered at x.)

Setting

Formalizing the problem:

- Neural network $f(W, x)$.
- Adversarial attack (PGD, FGSM, etc.) $\mathcal{A}(W, x) = x' \in \mathcal{B}(x)$
  ($\mathcal{B}(x)$ is the allowed perturbation set e.g. ℓ_2 or ℓ_∞ ball centered at x.)
- Adversarial training directly aims to minimize the *surrogate loss*

$$L_{\mathcal{A}}(W) = \frac{1}{n}\sum_{i=1}^{n} loss(f(W, \mathcal{A}(W, x_i)), y_i),$$

that is, the loss evaluated at the perturbed data generated by $\mathcal{A}$.

Setting

Formalizing the problem:

- Neural network $f(W, x)$.
- Adversarial attack (PGD, FGSM, etc.) $\mathcal{A}(W, x) = x' \in \mathcal{B}(x)$
  ($\mathcal{B}(x)$ is the allowed perturbation set e.g. ℓ_2 or ℓ_∞ ball centered at x.)
- Adversarial training directly aims to minimize the *surrogate loss*

$$L_{\mathcal{A}}(W) = \frac{1}{n}\sum_{i=1}^{n} loss(f(W, \mathcal{A}(W, x_i)), y_i),$$

  that is, the loss evaluated at the perturbed data generated by $\mathcal{A}$.
- While the true *robust loss* is

$$L_*(W) = \frac{1}{n}\sum_{i=1}^{n} \max_{x_i' \in \mathcal{B}(x_i)} loss(f(W, x_i'), y_i).$$

Setting (cont.)

- Fully-connected ReLU network, input dimension d, H hidden layers with width m.

Setting (cont.)

- Fully-connected ReLU network, input dimension d, H hidden layers with width m.
- Due to technical issues, we slightly modify the algorithm to *projected* adversarial training on a local region around initialization

$$B(R) = \left\{ W : \left\| W^{(h)} - W_0^{(h)} \right\|_F \leq \frac{R}{\sqrt{m}}, h = 1, \cdots, H \right\}.$$

Main Result

Theorem (Bounding the surrogate loss with the optimal robust loss)

Suppose $m \geq \text{poly}(R, H, d, 1/\epsilon)$. *With suitable assumptions and some* T *steps of training, we achieve*

$$\min_{t=1,\cdots,T} L_{\mathcal{A}}(W_t) \leq \min_{W \in B(R)} L_*(W) + \epsilon.$$

Corollary

Assume the network has approximation power $\min_{W \in B(R)} L_*(W) \leq \epsilon$, *then* $\min_{t=1,\cdots,T} L_{\mathcal{A}}(W_t) \leq 2\epsilon$.

Additional results

- For two-layer networks we derive a complete approximation result using random feature analysis.

Additional results

- For two-layer networks we derive a complete approximation result using random feature analysis.
- For two-layer networks, we derive a similar result without the need of projection.

Additional results

- For two-layer networks we derive a complete approximation result using random feature analysis.
- For two-layer networks, we derive a similar result without the need of projection.
- Why wide networks? We also derive an auxiliary VC-dimension result that implies achieving adversarial robustness requires more model capacity, e.g. width.

Thank you!

Welcome to our poster #115 for details and discussions!

Contact

Ruiqi Gao (grq@pku.edu.cn) and Tianle Cai (caitianle1998@pku.edu.cn) are applying for Ph.D. this year!

Please contact if you are interested!