Reinforcing Adversarial Robustness using Model Confidence Induced by Adversarial Training

Xi Wu

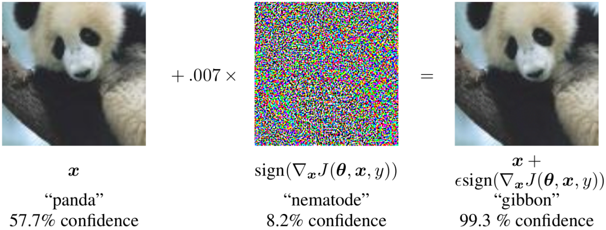
xiwu@cs.wisc.edu

Joint work with Uyeong Jang, Jiefeng Chen, Lingjiao Chen, and Somesh Jha

July 19, 2018

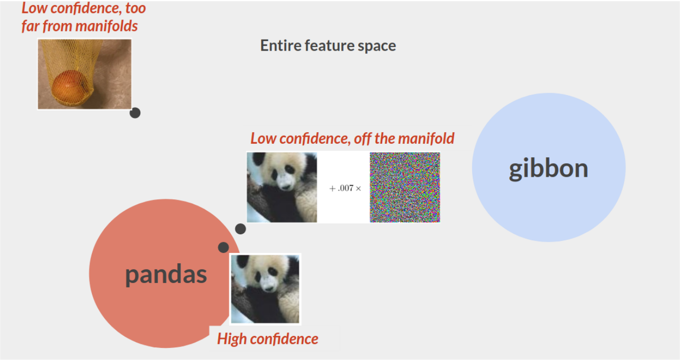
Entirely wrong behavior of confidence

- Small perturbations can cause highly confident but wrong predictions.
- An example from (Goodfellows, Shlens, and Szegedy, ICLR 2015), on a naturally trained neural network:



A better behavior

- *Low confidence if the model "does not learn/know it."*
- An intuitively good model for classifying pandas and gibbons (disks give natural data manifolds)

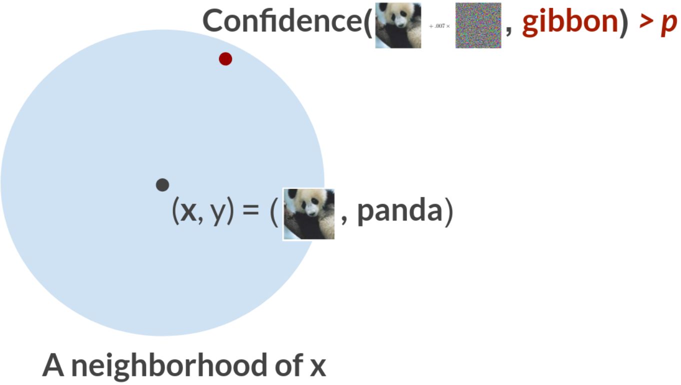


Main contributions of this work

- In a precise formal sense, adversarial training by (Madry et al., ICLR 2017) gives better behavior of model confidence for points near the data distribution.
- The better behavior of model confidence induced by adversarial training can be used to improve adversarial robustness.

Defining good behaviors of confidence (1/2)

Intuition: *Confident predictions of different classes should be well separated.*
A bad $(\mathbf{x}, y) \sim \mathcal{D}$ with poor confidence separation:



Defining good behaviors of confidence (2/2)

- $\mathcal{D}$: Data generating distribution; $d(\cdot,\cdot)$: A distance metric; $p, q \in [0,1]$, $\delta \geq 0$.
- Bad event (Neighborhood has p-confident wrong predictions):

$$\mathcal{B} = \{\exists y' \neq y, \mathbf{x}' \in N(\mathbf{x}, \delta), F_\theta(\mathbf{x}')_{y'} \geq p\}$$

- F is said to have (p, q, δ)-separation if

$$\Pr_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathcal{B} \right] \leq q.$$

Adversarial Training by Madry et al.

Adversarial training formulation of Madry et al.:

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

Theorem (Informal, this work)

For a large family of loss functions L, models trained as above achieve good (p, q, δ)-separation, where as $p \to 1$, $q \to 0$.

Empirical results (summary)

- We generate high-confidence attacks in order to bypass confidence-based defenses (as well as gradient-masking effect).
- Finding 1: Confidence of models trained using Madry et al.'s objective behave much better than their natural counterparts.
- Finding 2: A simple "nearest neighbor search" based on confidence corrects $20\% \sim 25\%$ targeted adversarial examples that fool the baseline model of Madry et al.
- Finding 3: For $> 98\%$ of test instances, correct label can be found in two neighbors with highest confidences.

Questions?

Please come to our poster session if you want to know more details!