CIS 7000-005: Debugging Data & Models Fall 2022

Robust Learning — October 20

Prof. Eric Wong

When you know what your model failures are, how do you fix them? One way is to retrain the model to be robust to these failures, an area known as robust learning.

- How to do adversarial training (robustness in the worst case)?
- How to do out of distribution robustness (robustness in the average case)?
- What are the empirical and guaranteed approaches, and what are the trade-offs between them?

1 Training for worst case robustness

1.1 Adversarial robustness

Robustness to adversarial examples has a checkered history. A lot of methods were proposed that simply didn't work. A brief history:

- In 2014 lots of attention brought to adversarial examples
- ICLR 2018 - 9 accepted papers on empirical adversarial robustness. 7 of the 9 were broken by Athalye et al 2018 before the conference even happened. There were also 2 certified defenses.
- January 2019 - white paper on adaptive attacks by Nicholas Carlini on how to really evaluate
- NeurIPS 2020 - 13 published papers on empirical adversarial robustness at ICLR, ICML, and NeurIPS are broken again by Tramer et al. 2020.
- Today - everyone uses some variant of adversarial training, the one defense not broken at ICLR 2018.

The concept behind adversarial training is simple. At a high level, we solve a minmax optimization problem:

$$\min_{\theta} \max_{\|\delta\| \leq \epsilon} \ell(f_\theta(x+\delta), y)$$

At each step, instead of training on the standard loss, we instead calculate an adversarial example for the current model and train on that loss instead. In practice we do this by running an adversarial attack, such as FGSM or its multi-step cousin PGD.

- The attack/defense landscape is assymetric. The attack that you use during training needs to be "strong enough" but does not have to be too strong. In contrast, at evaluation time you have to use a very strong attack.

- In fact it's possible to use FGSM adversarial training (one step), which is just 2x as long as standard training. The key is that a strong enough of an attack avoids a behavior known as *catastrophic overfitting*, but there are other ways to mitigate this as well.
- There is also *robust overfitt*, where adversarial robustness actually overfits if you train for too long even when standard deep learning does not.

1.2 Provable robustness

An alternative approach is to use provable guarantees to prove that no adversarial exists in training. At the core, this boils down to minimizing an upper bound on the adversarial loss:

$$\min_{\theta} \max_{\|\delta\| \leq \epsilon} \ell(f_\theta(x+\delta), y) \leq \min_{\theta} L(f_\theta, x, y, \epsilon)$$

These bounds can be gotten quickly with interval bound propagation or linear bounds based on linear programming or duality.

- Empirical defenses generally are faster and perform better.
- Provable defenses so far always take a hit to performance, and require training to get nonvacuous guarantees.

Randomized smoothing One other way to get provable robustness is to smooth a model over noise. This works for e.g. Gaussian smoothing to get ℓ_2 robustness. The way this works is to

- Sample a lot of ablations with random noise added
- Take a majority vote
- Check that the winning margin is large enough to hold with high probability

2 References