Game Theory and ML

Start Recording!

Lecture 5: Adversarial Examples

References to read for this course:

1. Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).

2. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014)

Last Time

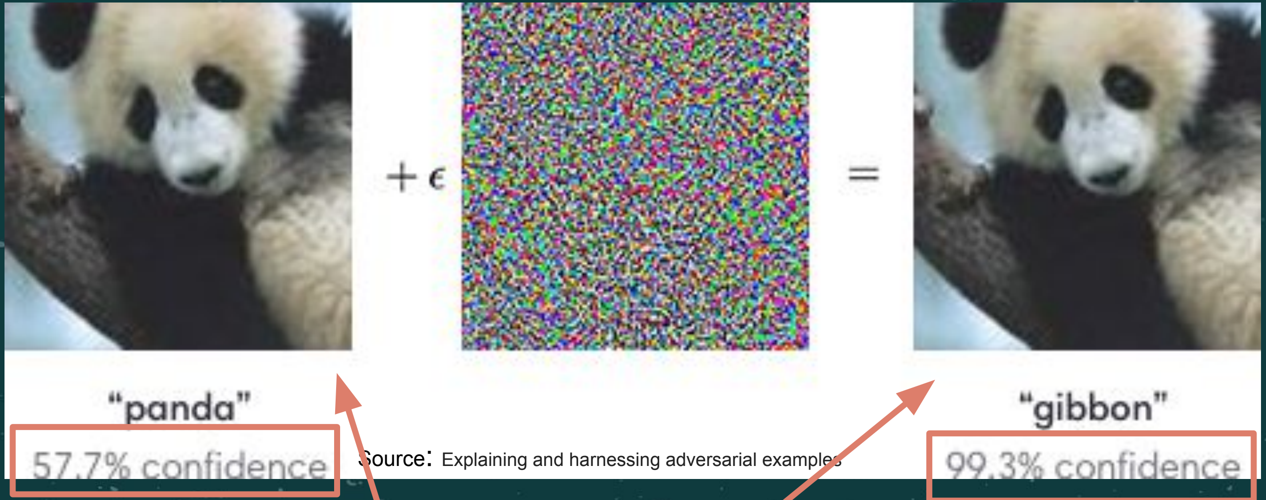
Empirical Risk Minimization:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

So what we have is a classifier that is good on the train set: (x_i, y_i)

Question What about what is close to the train and test sets?

Adversarial Examples in One Picture



Source: Explaining and harnessing adversarial examples

$$\|x - x'\| \leq \epsilon$$

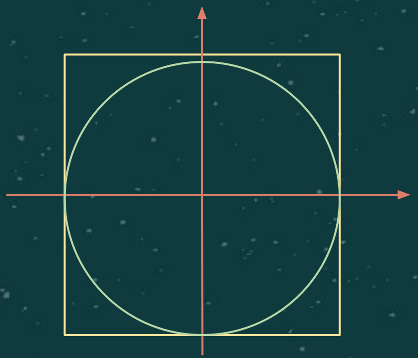
Adversarial Examples in One Picture

$$\|x - x'\| \leq \epsilon$$

Any meaningful norm

- Examples: L_2 or L_∞ norms.
- Beyond that anything that says

  'Two images are close'



How to find the Best attack?

Has to stay close to x!!!

$$\max_{x'} \ell(f(x'), y))$$

Such that $\|x - x'\| \leq \epsilon$

Adversarial Examples as an Optimization Problem

$$x' \in \arg\max_{x' \in \mathcal{X}} \ell(f_t(x'), y)\,, \quad \text{s.t.} \quad d(x, x') \leq \epsilon\,.$$

Optimization Problem

Usually L_p norm.

- f : function to attack.
- x : input datapoint.
- x’ : adversarial example.
- y : true label.
- ℓ : loss function.

We Need to know the function to optimize

Threat Model

$$x' \in \arg\max_{x' \in \mathcal{X}} \ell(f_t(x'), y)\,, \quad \text{s.t.} \quad d(x, x') \leq \epsilon\,.$$

Optimization

- Threat model:
  What do we assume the attacker has access to.
  (i.e. what is the threat)

- White Box threat model: Access to the gradients of f
- Black Box threat model: Access to the values of f
- Practical Black Box (see today's presentation)
- NoBox threat model: (see following lectures)

Threat Model

$$x' \in \arg\max_{x' \in \mathcal{X}} \ell(f_t(x'), y)\,, \quad \text{s.t.} \quad d(x, x') \leq \epsilon\,.$$

Optimization

- Threat model:
  What do we assume the attacker has access to.
  (i.e. what is the threat)

- White Box threat model: Access to the gradients of f
- Black Box threat model: Access to the values of f
- Practical Black Box (see today's presentation)
- NoBox threat model: (see following lectures)

White Box Threat Model

- Idea: Use (projected) gradient **ascent** to solve this:

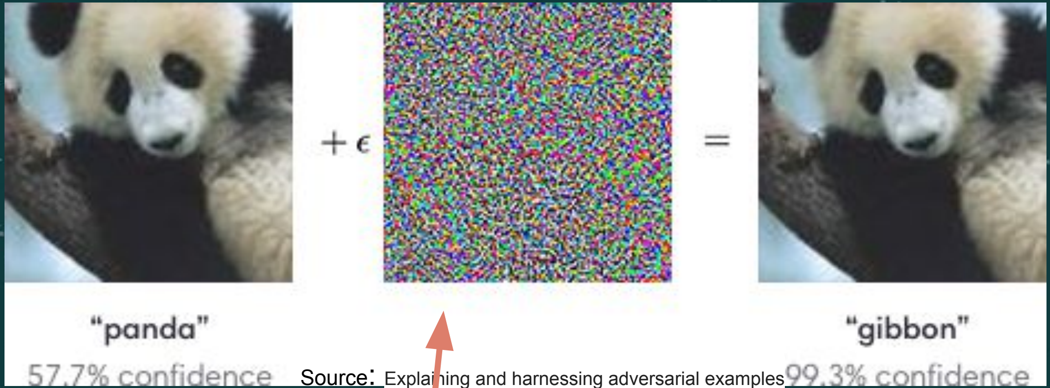
$$x' \in \arg\max_{x' \in \mathcal{X}} \ell(f_t(x'), y)\,, \quad \text{s.t.} \quad d(x, x') \leq \epsilon\,.$$

- Idea 2: Use a gradient method that correspond to the geometry of the constraint:

$$||x - x'|| \leq \epsilon$$

- Idea 3: do we need more than 1 step?

Original attack

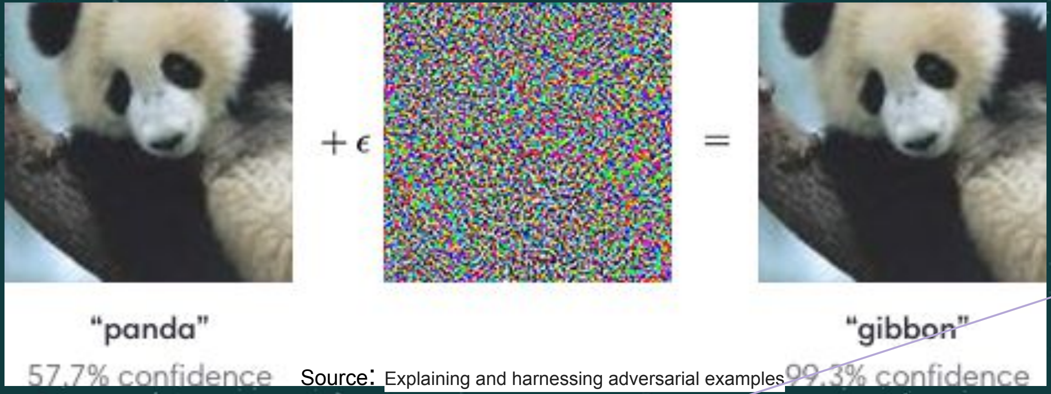


Source: Explaining and harnessing adversarial examples

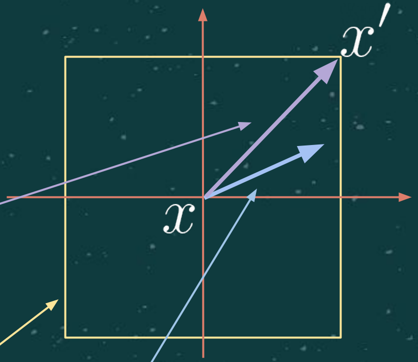
$$\cancel{\nabla_x \ell(f(x), y)})$$

$$\text{sign}(\nabla_x \ell(f(x), y)))$$

Original attack



Source: Explaining and harnessing adversarial examples



$$\epsilon \cdot \operatorname{sign}(\nabla_x \ell(f(x), y)) = \max_{\|p\|_\infty \leq \epsilon} p^\top \nabla_x \ell(f(x), y)$$

- 1 step of Steepest descent with L_∞ constraints.

Fancier attacks

- Several steps of steepest descent. [Madry et al, 2017]
- Add momentum [Dong et al 2018]
- When several steps.. Be careful of the constraints:

$$x' \in [0,1]^d \quad \text{and} \quad ||x - x'|| \leq \epsilon$$

Black Box attack

- Idea: query ℓ around x to get an approximation of the gradient.

$$\frac{\ell(f(x)) - \ell(f(x+\delta e_i))}{\delta} \approx [\nabla \ell(f(x))]_i$$

Related to zero-th order optimization

(will not be the topic of this course)

- See [Siddhant et al. 2020] for a survey of Black Box Attacks.

Defenses

Ideas to be robusts againsts such Adv Attacks:

1. Gradient Masking (now)
2. Preprocessing of the input
3. Adversarial Training (Next Lecture)
4. Many more…[Prakash et al. 2018], [Liao et al. 2018], [Schott, Lukas, et al. 2018] (Open research direction)
   (see https://www.robust-ml.org/defenses/)

Useful References:

1. Szegedy, Christian, et al. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013).

2. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014)
3. https://openai.com/blog/adversarial-example-research/
4. Papernot, Nicolas, et al. "Practical black-box attacks against machine learning." *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017. (**Presentation on that paper)**
5. Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
6. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. Sixth International Conference on Learning Representations (ICLR), 2017

Standard things to know

- Change the loss [Carlini, Wagner 2016].
- Targeted adversarial attacks:

Other label

$$x' \in \arg \min_{\|x - x'\| \leq \epsilon} \ell(f(x'), y')$$