Evidence for adversarial robustness through randomization: a game theoretic perspective?

Toward certified defenses to adversarial example attacks

Rafael PINOT

Paris Dauphine University
CEA LIST Institute
Lamsade day 2019.

Overview

I. Introduction to Supervised Learning & Neural Networks.

II. Adversarial attack & Defense through randomization.

III. A Game theoretic perspective.

I. Introduction to Supervised Learning & Neural Networks

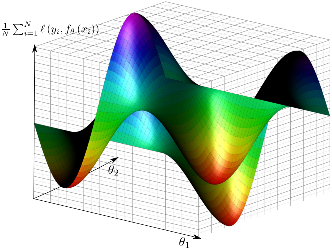
What is Supervised Learning

$f(x_i) = y_i$	
x_1 	$y_1 =$ "*dog*"
x_2 	$y_2 =$ "*panda*"
x_n 	$y_n =$ "*cat*"

- Given a set of *n* **training examples** $\{(x_1, y_1), ..., (x_n, y_n)\} \sim D$.
- **Assumption:** there exists a mapping f matching any vector to its label.

Learning algorithm goal: Approximate f by a parametrized function f_θ.

Supervised Learning Algorithms



- To measure how well f_θ fits f, owe use a **loss function** $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.
- Find the parameter θ that minimize the **generalization error**

$$\boxed{\mathbb{E}_{(x,y)\sim D}\left[\ell\left(y, f_\theta\left(x\right)\right)\right]}$$

The standard method to find θ is the **empirical risk minimization (ERM)**:

$$\hat{\theta}_{ERM} := \text{argmin}_{\theta\in\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(y_i, f_\theta\left(x_i\right)\right)\right] \text{ recall: } y_i = f(x_i)$$

Neural networks

A **neural network** is a directed and weighted graph, modeling the structure of a **dynamic system**. A neural network is analytically described by list of function compositions.

A **Feed forward neural network** of N layers is defined as follows:

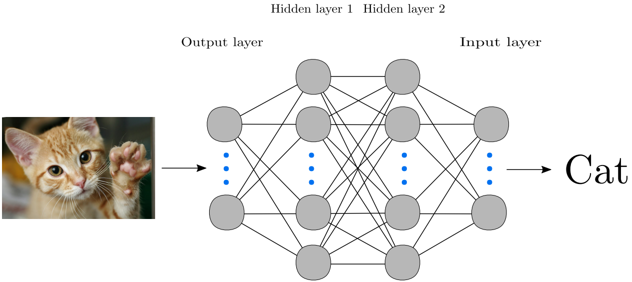
$$f_{\hat{\theta}_{ERM}} := \phi^{(N)} \circ \phi^{(N-1)} \circ \cdots \circ \phi^{(1)}(x)$$

Where for any i, $\phi^{(i)} := z \mapsto \sigma(W_i z + b_i)$, $b_i \in \mathbb{R}^m$, $W_i \in \mathcal{M}_{\mathbb{R}}(m, n)$ (n size of z), and σ some non linear (activation) function.

Feed forward networks, as well as some other specific types of network are said to be **universal approximators** [Cybenko, 1989].

Deep neural networks

Deep neural networks (large and complex networks) has recently proven outstanding results especially in **image classification**.



No free lunch:

1) (Deep) Neural networks lack theoretical guarantees.
2) The model is often over-parametrized, which can lead to over-fitting, or to other **flaws in the classification task** (e.g adversarial examples).

Attacking Models & Defense through Randomization.

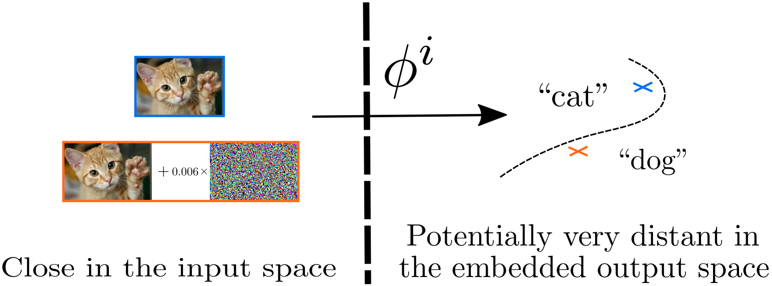
Adversarial examples

An **adversarial attack** refers to a small, imperceptible change of an input maliciously designed to fool the result of a machine learning algorithm.

Since the seminal work of [Szegedy et al., 2014] exhibiting this intriguing phenomenon in the context of deep learning, numerous attack methods have been designed (e.g. [Papernot et al., 2016, Carlini and Wagner, 2017]).

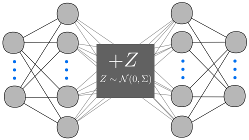
Geometric interpretation

Adversarial example: Neural networks do not preserve distances between images. adversaries take advantage of it to find adversarial examples.



How to defend? A learning algorithm should be robust to adversarial examples, if it has a local (small ball around each image) isometric property.

Defense by randomization



- [Dhillon et al., 2018]: Sampling parameters from a probability measure.
- [Pinot et al., 2019]: Add well selected random noise to parameters.

For a **Feedforward network**, we modify one/several functions:

$$\tilde{f}_{\hat{\theta}_{ERM}}(x) = \phi^{(N)} \circ \cdots \circ \tilde{\phi}^{(i)} \circ \cdots \circ \phi^{(1)}(x)$$

Where$\tilde{\phi}^{(i)}(z) = \sigma(W_i z + b_i), and (W_i, b_i) \sim \mathcal{N}(0, \Sigma)$.

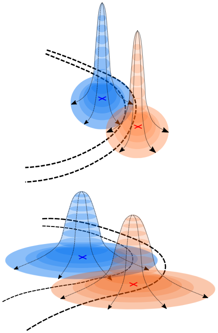
How does this sampling work?

Several possible interpretations:

1) Robust optimization: Noise helps locally smoothing the network.

2) Geometrical: Noise pushes the decision boundary/makes it "probabilistic".

3) **Game theory:** Equilibrium is achieved using mixed strategies.



A game theoretic perspective

Strategic game:

A strategic game G is defined by:

- A set I of N players.
- A set S^i of strategies (one per player).
- An application $g : S = \prod_{i \in [N]} S^i \mapsto \mathbb{R}^N$ where each component is g^i represents the utility function of player i over S.

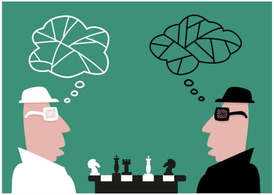
Goal: Strategic games are one-shot games, where everybody plays its move simultaneously. One may want to find out if there exists an equilibrium?

Equilibrium: An equilibrium is a strategy $s \in S$ where no player can increase her utility by deviating alone from s.

Zero sum game & mixed strategies

Zero-sum game:

- There is only 2 players.
- For any $s \in S$, one has $g^1(s) = -g^2(s)$
- Is characterized by (S^1, S^2, g^1).



Mixed strategies: Under mild assumptions on S^1, S^2, and g^1 one equilibrium can be found in the mixed extension of the game.
This extension is as follows: $(\Delta S^1, \Delta S^2, \mathbb{E}[g^1])$ where $\Delta(\bullet)$ denotes the set of probability measures over $\bullet$.

Adversarial attacks as a game

Goal of the model (defender): find θ that minimize the loss over all possible (potentially adversarial) examples.

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim D} \left[\ell \left(y, f_\theta \left(x \right) \right) \right]$$

Vs

Goal of the adversary: find the best way (best function) to compute a perturbation that fools f_θ for any image x from the ground-truth distribution.

$$\max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim D} \left[\ell \left(y, f_\theta \left(x + g\left(x \right) \right) \right) \right]$$

A Zero-sum game

Can be seen as a **Zero-sum** game with the following objective function:

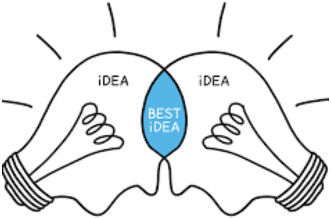
$$\boxed{\mathbb{E}_{(x,y)\sim D}\left[\ell\left(y, f_\theta\left(x + g\left(x\right)\right)\right)\right]}$$

- From classical game theory, it makes sense to use mixed strategies, i.e players should sample from Θ and $\mathcal{G}$, according to some distribution.
- The solution we gave earlier is a mixed strategy. Is it an optimal one?

We can use game theoretical arguments to justify procedures from machine learning, and to study them!

Conclusion

- Treating the problem of adversarial example can be **both** considered as a game theoretic and a machine learning problem.
- Both seem to converge to the same conclusion: **randomization matters**.



Conclusion: Pole 1, 2, and 3 should work together on this issue ;-).

Carlini, N. and Wagner, D. (2017).
Towards evaluating the robustness of neural networks.
In 2017 IEEE Symposium on Security and Privacy (SP), pages 39–57. IEEE.

Cybenko, G. (1989).
Approximation by superpositions of a sigmoidal function.
Mathematics of control, signals and systems, 2(4):303–314.

Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., and Anandkumar, A. (2018).
Stochastic activation pruning for robust adversarial defense.
CoRR, abs/1803.01442.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016).

The limitations of deep learning in adversarial settings.
In Security and Privacy (EuroS&P), 2016 IEEE European Symposium on, pages 372–387. IEEE.

Pinot, R., Meunier, L., Araujo, A., Kashima, H., Yger, F., Gouy-Pailler, C., and Atif, J. (2019).
Theoretical evidence for adversarial robustness through randomization: the case of the exponential family.
CoRR, abs/1902.01148.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014).
Intriguing properties of neural networks.
In International Conference on Learning Representations.