Toward certified defenses to adversarial example attacks trough randomization

Rafael PINOT & Cédric GOUY-PAILLER & Jamal ATIF

CEA LIST Institute
Paris Dauphine University.

Overview

I. Introduction to Supervised Learning & Neural Networks.

II. Adversarial example attacks.

III. Defense methods & randomization.

I. Introduction to Supervised Learning & Neural Networks

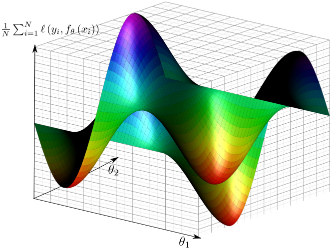
What is Supervised Learning

$f(x_i) = y_i$	
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, we use a **loss function** $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$.
- Find the parameter θ that minimizes 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:

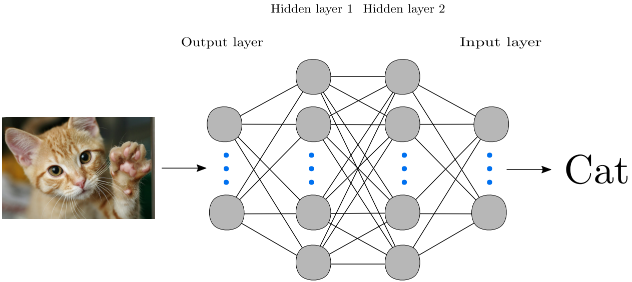
$$F := f_{\hat{\theta}_{ERM}} = \phi^{(N)} \circ \phi^{(N-1)} \circ \cdots \circ \phi^{(1)}(x)$$

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).

Adversarial example attacks

Intriguing properties of neural networks

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 [Biggio et al., 2013] 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]).

Crafting an adversarial attack

Let $\mathcal{X}$ and $\mathcal{Y}$ be respectively the image and the label spaces. Let us also consider (x, y) a labeled image. **To craft the adversarial example:**

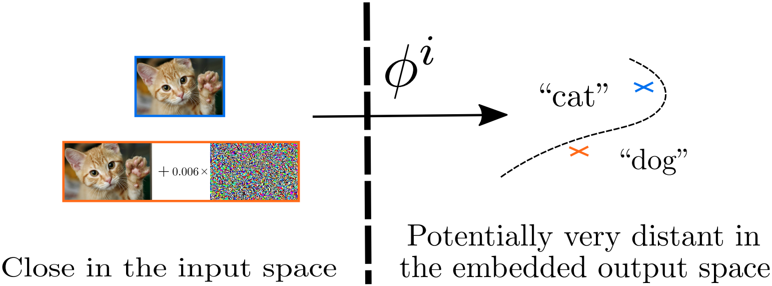
- The adversary should solve $\min\limits_{F(x+\tau)\neq y} ||\tau||$ which is hard.
- This is relaxed to $\min c||\tau|| - \ell(y, F(x+\tau))$ with $c > 0$.
- One can simply take $x^{\text{adv}} = x + \gamma \frac{\nabla_x \ell(y,F(x+\tau))}{||\nabla_x \ell(y,F(x+\tau))||}$ (small enough γ).

This very simple attack make the classifier's accuracy **drop drastically**.

Some (not much) more sophisticated attacks make the **accuracy drop to 0%**.

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 methods & randomization

Current state-of-the-art: Adversarial training

- At every step of the learning procedure, for each image, augment the batch with corresponding adversarial example (see [Madry et al., 2018]).
- Gives an 'ok' defense against adversarial examples (here CIFAR10).
- Adversarial training is **computationally costly**.
- Provides no theoretical analysis, **hence no worst case behavior**.

Attack	**Steps**	Madry et al.
-	-	0.873
ℓ_∞ – PGD	20	0.456
ℓ_2 – C&W	30	0.468

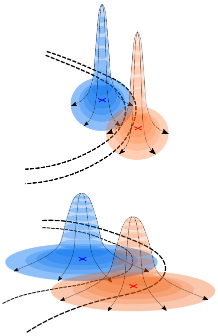
An other technique: randomization

- Randomization is massively studied in a lot of domains.
- Provides theoretical background/rationale of the defense mechanism.
- In some cases, it provides **theoretical results on the worst case scenario**.
- In some cases, it can be **computationally efficient**.

Interpretation randomization

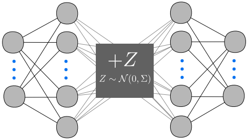
Several possible interpretations and techniques:

- Robust optimization: Noise helps locally smoothing the network.
- Data augmentation: Noise helps the network minimize the generalisation error.
- **Topological:** Change the output space to be a space of probability distributions.
- Game theory: there is no pure Nash equilibrium $\implies$ one needs a mixed strategy.



Our point of view: Topological

Recent Works [Li et al., 2019, Cohen et al., 2019, Pinot et al., 2019] propose to inject noise at a given layer of the network **at inference**.



Formally: for a Feedforward network, we have

$$\tilde{F}_\epsilon(x) = \phi^{(N)}_{W_N, b_N} \circ \cdots \circ \tilde{\phi}^{(i)}_{W_i, b_i} \circ \cdots \circ \phi^{(1)}_{W_1, b_1}(x)$$

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

Then one can use the expectation over transformations as a robust classifier:

$$F^{\text{rob}}(x) := \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Sigma)} \left(\tilde{F}_\epsilon(x) \right)$$

Geometrical interpretation

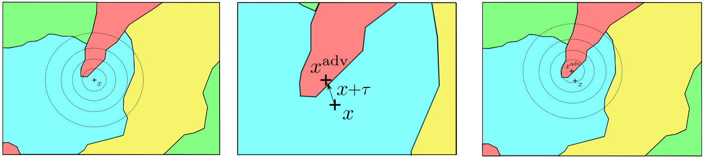


Figure inspired by [Cohen et al., 2019]

- Adding $\mathcal{N}(0, \sigma^2)$ to the natural image produces a probability **distribution on the regions** $\mathbb{P}[X \in \text{region}]$ with $X \sim \mathcal{N}(x, \sigma^2)$.
- Adding the same noise on x^{adv} produces almost the same distribution.
- Hence $F^{\mathrm{rob}}(x)$ and $F^{\mathrm{rob}}(x^{\mathrm{adv}})$ should give **similar results**.

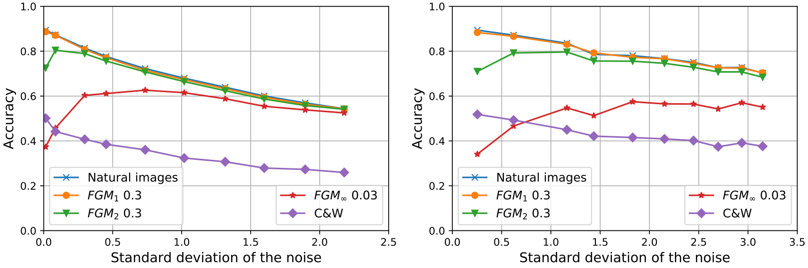
Formal validation

From [Cohen et al., 2019]:

Let $F^{\text{rob}}(x) := \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)}\left(\tilde{F}_\epsilon(x)\right)$ be the classifier at hand. $\exists \alpha^* > 0$ such that, for any $||\tau|| < \alpha^*$ one has $F^{\text{rob}}(x) = F^{\text{rob}}(x+\tau)$

- Noise injection gives a **worst case certificate**.
- We [Pinot et al., 2019] extended this work to any exponential family.
- Values of α^* are **still to small** for the methods to be fully robust.

Some numerical results



- **Trade-off** between robustness to attacks, and accuracy of the method.
- Best attacks remain **hard to mitigate**.

Take home message

- Adversarial examples are a burning issue and a big security breach.
- Randomization presents principled advantages over other defenses.
- Overall defense capabilities remain weak.
- Room for improvement both theoretically (bigger α^*) and experimentally (try more distributions, and more sophisticated randomized settings).

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013).
Evasion attacks against machine learning at test time.
In Joint European conference on machine learning and knowledge discovery in databases, pages 387–402. Springer.

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.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. (2019).
Certified adversarial robustness via randomized smoothing.
CoRR, abs/1902.02918.

Cybenko, G. (1989).

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

Li, B., Chen, C., Wang, W., and Carin, L. (2019).
Certified adversarial robustness with addition gaussian noise.
In Advances in Neural Information Processing Systems 32 (NeurIPS).

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018).
Towards deep learning models resistant to adversarial attacks.
In International Conference on Learning Representations.

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.
In Advances in Neural Information Processing Systems 32 (NeurIPS).