Project: Adversarial Attacks

How to build robust Neural Network.

Alexandre ARAUJO - Rafael PINOT - Geovani RIZK

Paris Dauphine, PSL
MILES research team.

Dauphine | PSL

About us

+ **Alexandre ARAUJO** Paris Dauphine University - Wavestone.
  - Deep Learning theory and application.
  - Linear algebra and compression algorithms.
  - Robustness to adversarial examples.

+ **Rafael PINOT** Paris Dauphine University - CEA Saclay.
  - Machine learning and information theory.
  - Probability and randomized algorithms.
  - Privacy and anonymization mechanisms.

+ **Geovani RIZK** Paris Dauphine University - Huawei.
  - Deep Learning applications and Generative methods.
  - Game theory and Bandit algorithms.
  - Robustness to adversarial examples.

Supervised Learning Algorithms

X					Y
$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,p-1}$	$x_{1,p}$	y_1
$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,p-1}$	$x_{2,p}$	y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{n-1,1}$	$x_{n-1,2}$	$\cdots$	$x_{n-1,p-1}$	$x_{n-1,p}$	y_{n-1}
$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,p-1}$	$x_{n,p}$	y_n

Given a set of n **training examples** $\{(x_1, y_1), ..., (x_n, y_n)\}$ where x_i is the feature vector of the i^{th} example, and y_i the is the corresponding label.

Assumption: there exists a function f matching any feature vector to its label.

A **learning algorithm** goal is to approximate f by a parametrized function f_θ.

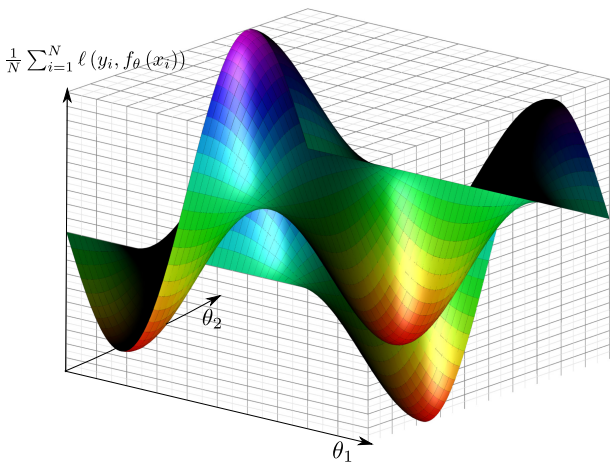
In order to measure how well the function fits, **a loss function** $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is defined. The standard method to learn the parameter θ is the **empirical risk minimization (ERM)**:

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

Model specification and consistency

The model is said to be **correctly specified** if there exists a parameter θ^* such that $f_{\theta^*} = f$. Otherwise, the model is said to be **misspecified**.

It is usually **hard to design correctly specified models in machine learning**. Hence we separate models into consistent and non consistent models.



A model is said to be **consistent** if $\hat{\theta}_{ERM}$ converges in probability to θ^*, as $n \to \infty$ (if the model is correctly specified, the average loss of f_θ is null).

When ℓ is differentiable, **in order to converge** to a (local of global) minima, one often uses **gradient based methods**.

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:

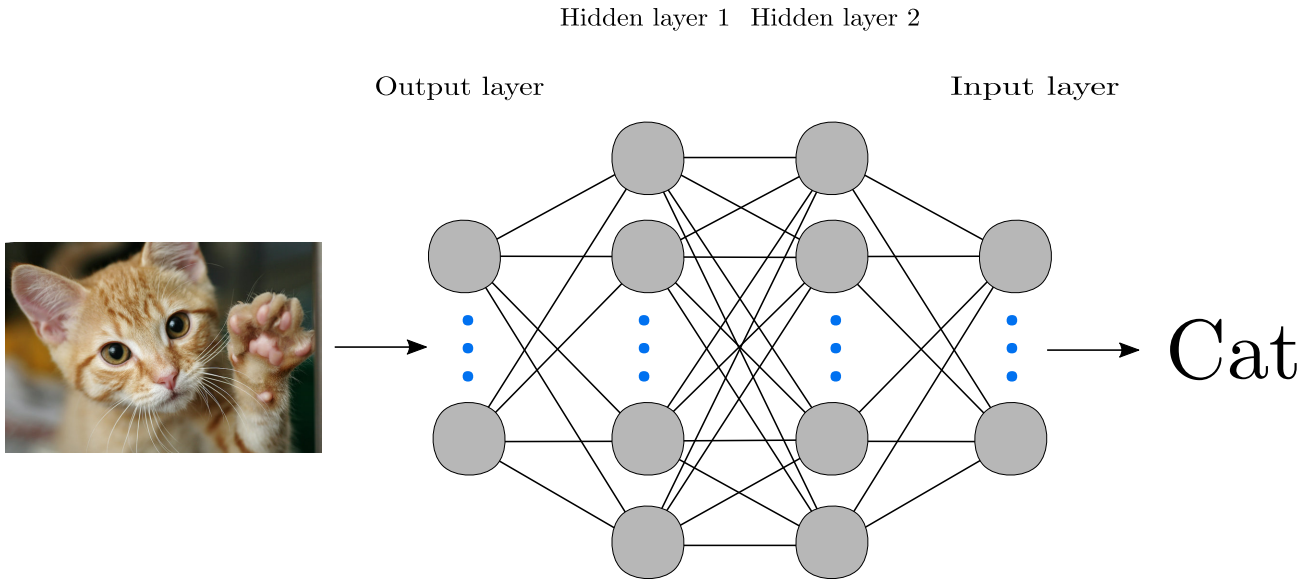
$$f_{\hat{\theta}_{ERM}} := F(x) = \phi^{(N)}_{W_N, b_N} \circ \phi^{(N-1)}_{W_N - 1, b_N - 1} \circ \cdots \circ \phi^{(1)}_{W_1, b_1}(x)$$

Where for any i, $\phi^{(i)}_{W_i, b_i} := z \mapsto \sigma(W_i z + b_i)$, $b_i \in \mathbb{R}^m$, $W_i \in \mathcal{M}_{\mathbb{R}}(m, n)$ (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) have recently proven outstanding results especially in **image classification**.



No free lunch:

1) Neural networks (especially deep neural networks) lack theoretical guarantees regarding specification and consistency of the model.
2) The model is often over-parametrized, which can lead to over-fitting, or to **major flaws in the classification task** (e.g adversarial examples).

Attacking a Neural Network.

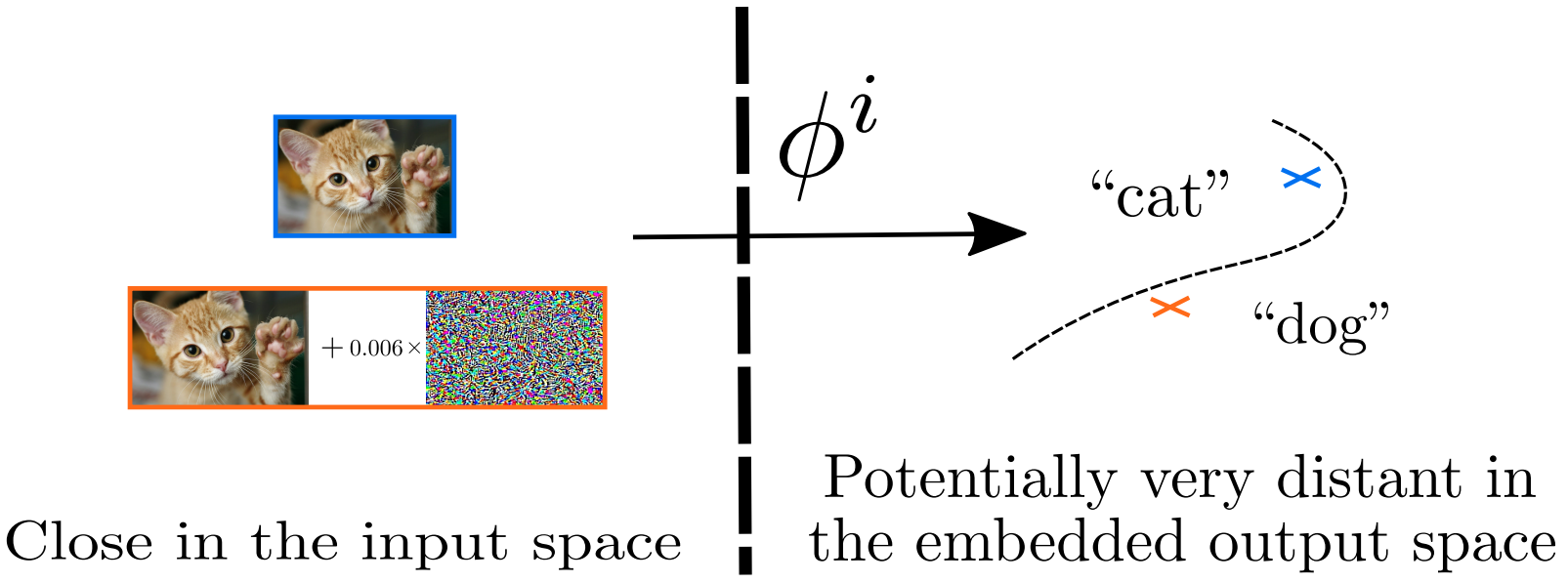
Intriguing properties of neural networks

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 [Szegedy et al. 2013] exhibiting this intriguing phenomenon in the context of deep learning, numerous attack methods have been designed (e.g. [Carlini et al. 2016, Madry et al.]).

Geometric interpretation

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

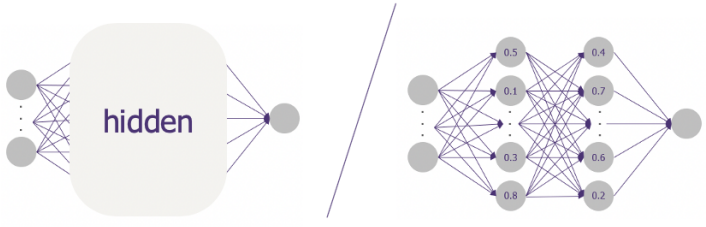


How to defend? A learning algorithm should be robust to adversarial examples, if it has a local (small ball around each image) isometric property.

Effective defense: Among the defenses, noise injection at training/inference time to the network is demonstrated great experimental performances.

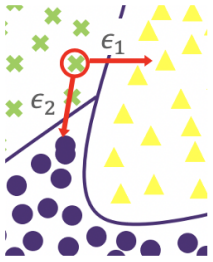
Different types of attacks

Black Box / White Box attacks



In white box attacks the attacker has access to the model's parameters, while in black box attacks the attacker has no access to these parameters, i.e., it uses a different model or no model at all to generate adversarial images with the hope that these will transfer to the target model.

Targeted / Untargeted attacks



The aim of non-targeted attacks is to reinforce the model to misclassify the adversarial image, while in the targeted attacks the attacker can get the input classified with a chosen specific target class, which is different from the true class.

Our goals in this project

- **Notebook 1** Introduction: Adversarial Attacks on a linear model.
- **Notebook 2** Introduction of FGSM and PGD attacks on Neural Network.
- **Notebook 3** Adversarial Training: How to build robust classifier.

Some introducing references

- **Szegedy et al. 2013:**
  Intriguing properties of neural networks.
- **Goodfellow et al. 2014:**
  Explaining and Harnessing Adversarial Examples.
- **Carlini et al. 2016:**
  Towards Evaluating the Robustness of Neural Networks.
- **Madry et al. 2017:**
  Towards Deep Learning Models Resistant to Adversarial Attacks.