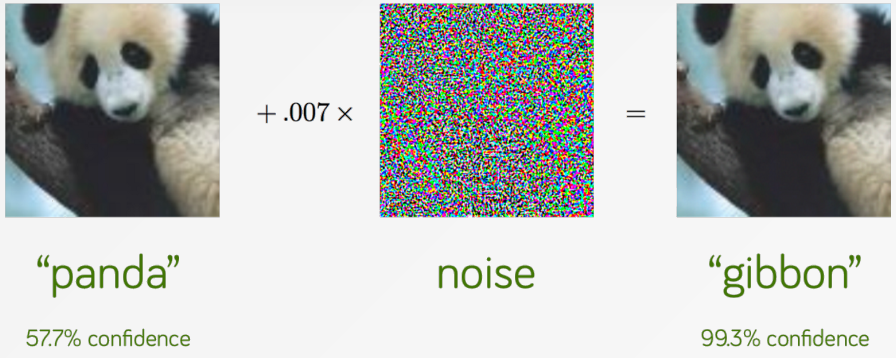
Certifying robustness using optimal transport

IRT: Franck Mamalet, Thibaut Boissin
IRIT: Mathieu Serrurier *, Louis Bethune *
IMT: Alberto Gonzalez Sanz *, Jean-Michel Loubes *

25/03/2022

* : Aniti's chair Robust & Fair Learning

One noise to fool them all



Deep Neural Networks are

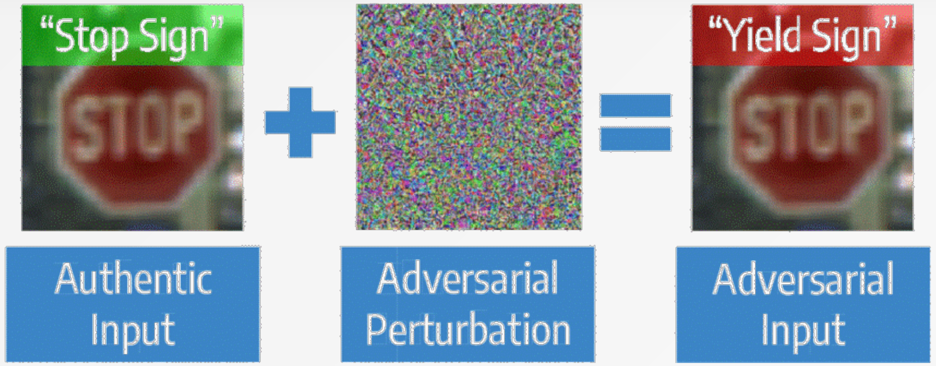
- Very Efficient
- Highly unstable
- Sensitive to small Perturbations

Difficult to Certify their behaviour for **high risk systems in industry**

Adversarial attacks:

$$\mathrm{adv}(f, x) = \underset{x',\ f(x) \neq f(x')}{\operatorname{argmin}} \|x - x'\|$$

One noise to fool them all



Deep Neural Networks are

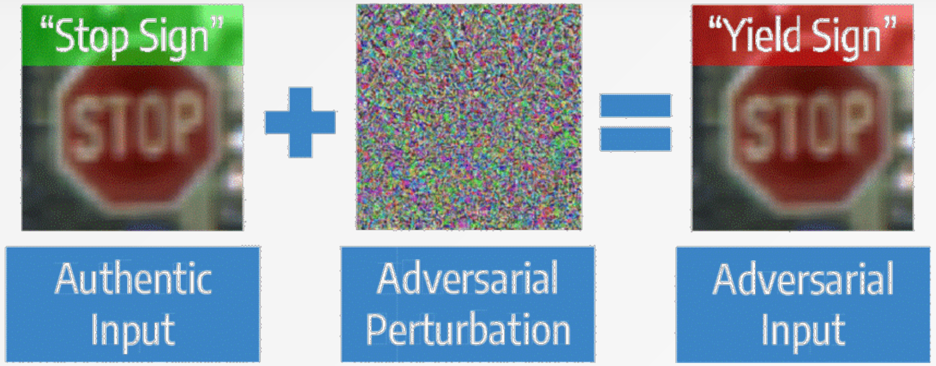
- Very Efficient
- Highly unstable
- Sensitive to small Perturbations

Difficult to Certify their behaviour for **high risk systems in industry**

Adversarial attacks:

$$\mathrm{adv}(\mathrm{f}, \mathrm{x}) = \underset{x',\ f(x) \neq f(x')}{\operatorname{argmin}} \|x - x'\|$$

One noise to fool them all



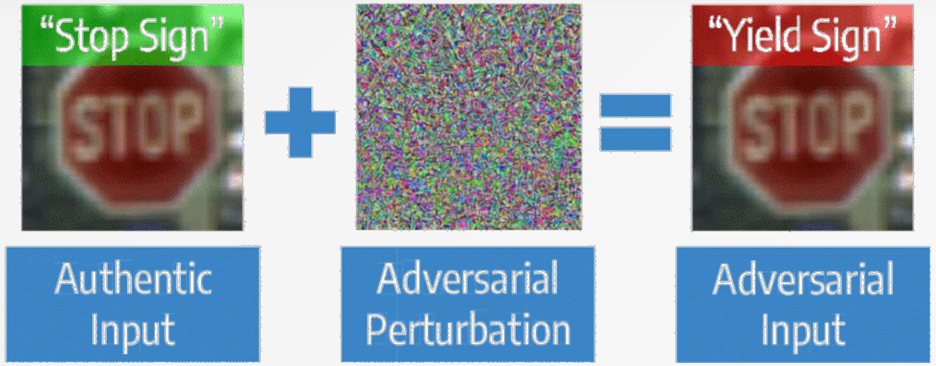
Deep Neural Networks are
- Very Efficient
- Highly unstable
- Sensitive to small Perturbations

Difficult to Certify their behaviour for **high risk systems in industry**

Adversarial attacks:

$$\mathrm{adv}(f, x) = \underset{x',\ f(x) \neq f(x')}{\operatorname{argmin}} \left\| x - x' \right\|$$

One noise to fool them all



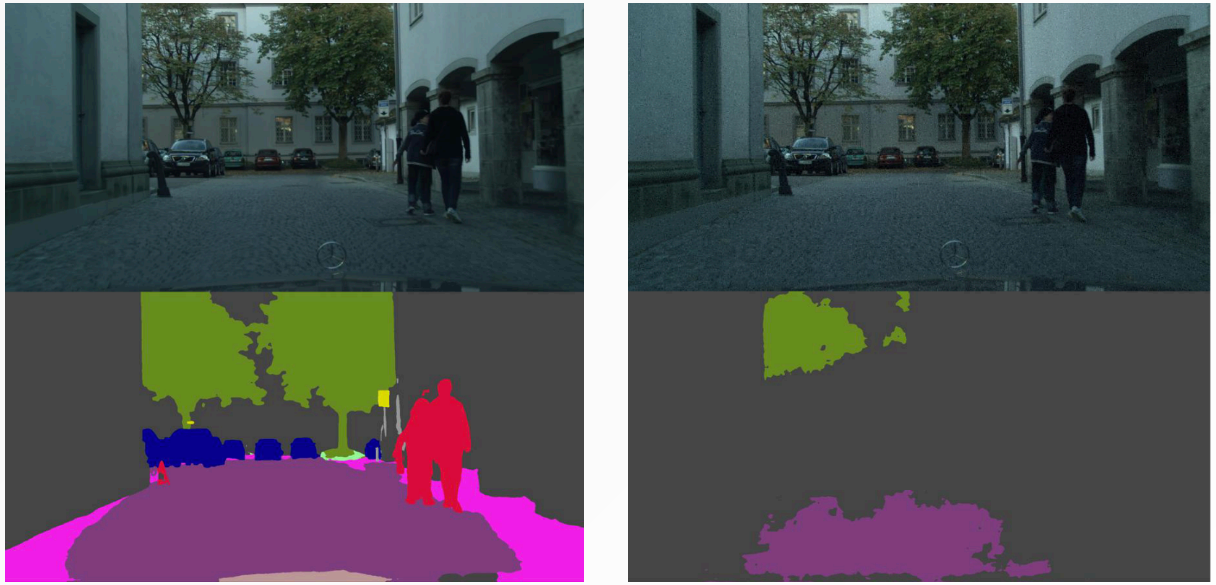
Deep Neural Networks are

- Very Efficient
- Highly unstable
- Sensitive to small Perturbations

Difficult to Certify their behaviour for **high risk systems in industry**

Adversarial attacks:

$$\mathrm{adv}(\mathrm{f}, \mathrm{x}) = \underset{x',\ f(x) \neq f(x')}{\operatorname{argmin}} \left\| x - x' \right\|$$



HINGE KANTOROVICH-RUBINSTEIN ROBUST CLASSIFIER (CVPR 2021)

Our contribution is a combination of

1. **Lipschitz property** constraint to ensure Robustness
2. Theory of **Optimal Transport** applied to classification

Resulting an

Optimal classifier with provable robustness guarantees

- <u>**Main Theorems**</u>

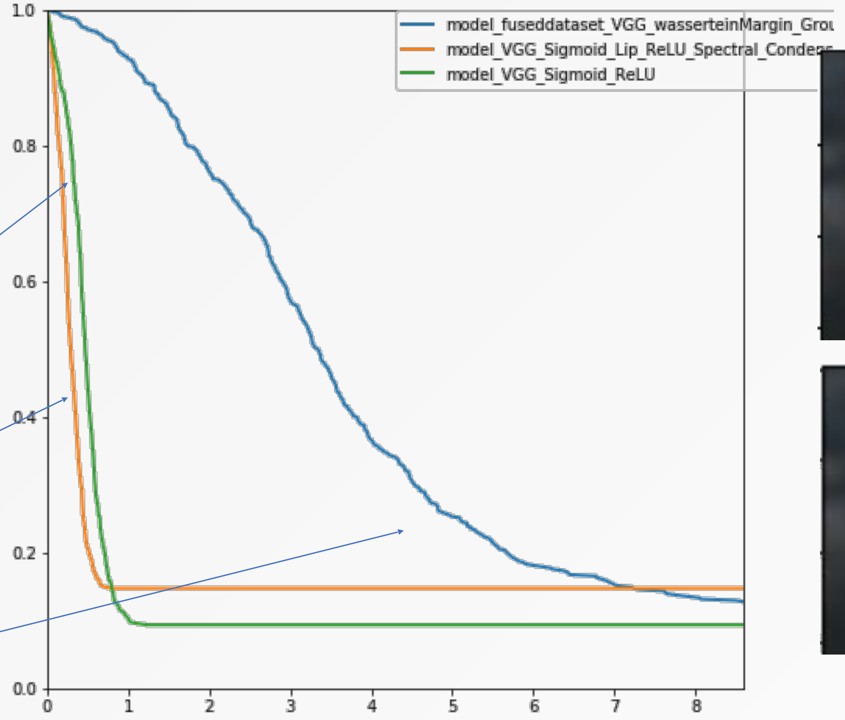
1. Existence and Uniqueness of the classifier
2. Provable Generalization guarantees and state of the art performance
3. Certifiable robustness

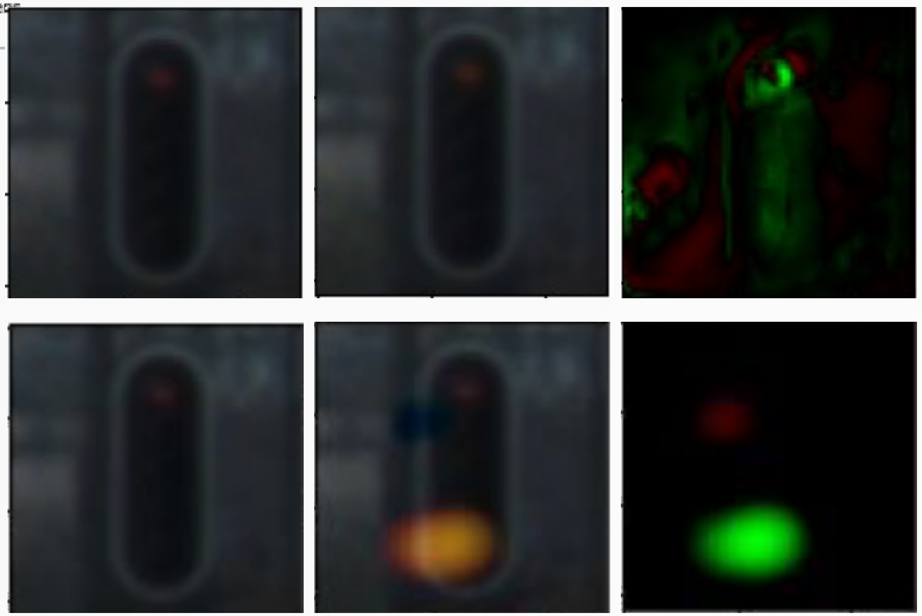
Hinge Kantorovich-Rubinstein Robust classifier (CVPR 2021)

Unconstrained network

Constrained network

Our network





Unconstrained network

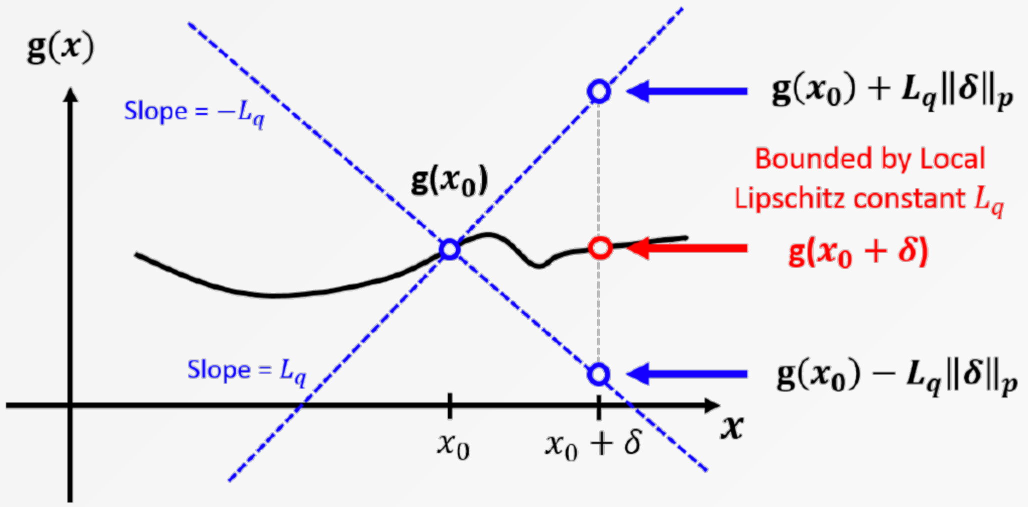
Our network

Lipschitz Property in Machine Learning

Lipschitz property of function enhances **robustness**

$$\| f(x) - f(y) \| \leq L_\star \| x - y \|$$

$$\| \nabla_x f \| \leq L_\star$$



$$\| x - y \| \leq \varepsilon \rightarrow \| f(x) - f(y) \| \leq L_\star \varepsilon \leq \delta$$

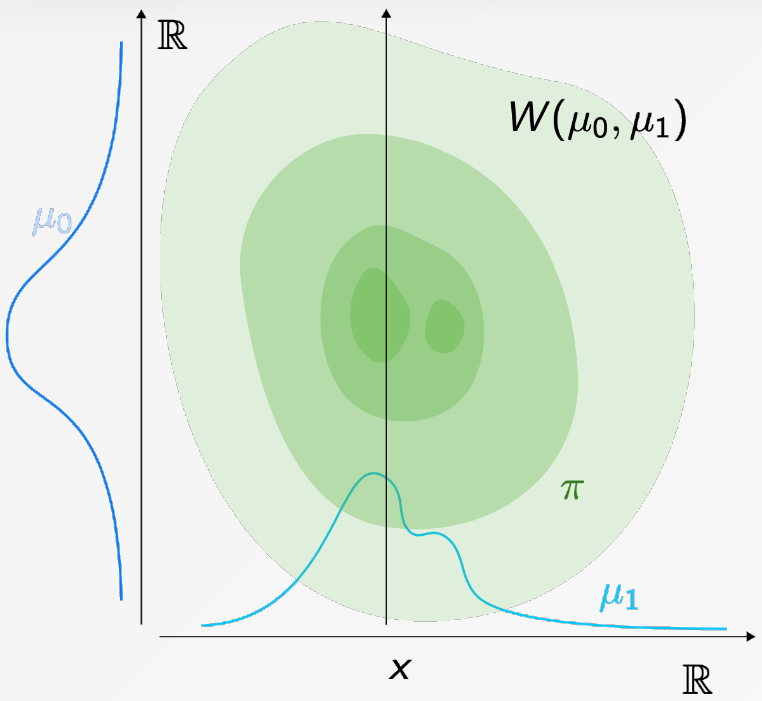
Deep Neural Networks suffer from **explosion of the Lipschitz constant** due to

- Structure : the deeper, the less control. 'p layers' $f(x) = f_1 \circ f_2 \circ \dots f_p(x)$

$$\| f(x) - f(y) \| \leq \underbrace{L_1 \times \dots \times L_p}_{L_\star} \| x - y \|$$

- Minimizing cross-entropy for better accuracy is adverse to Lipschitz smoothness.

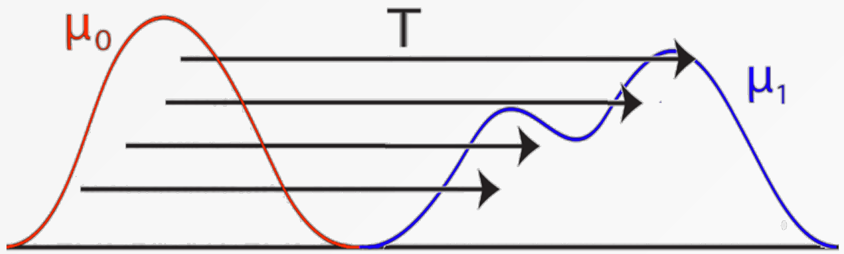
Optimal Transport Theory from Monge to Kantorovich



$$\mathscr{W}_c(\mu_0, \mu_1) = \underset{\pi \in \Pi(\mu_0, \mu_1)}{argmin} \int c(x, y) d\pi(x, y)$$

$$= \underset{T, T(X) \sim \mu_1}{argmin} \int c(x, T(x)) d\mu_0(x)$$

1-Lipschitz functions are related to Optimal Transport for $c(x, y) = \| x - y \|$

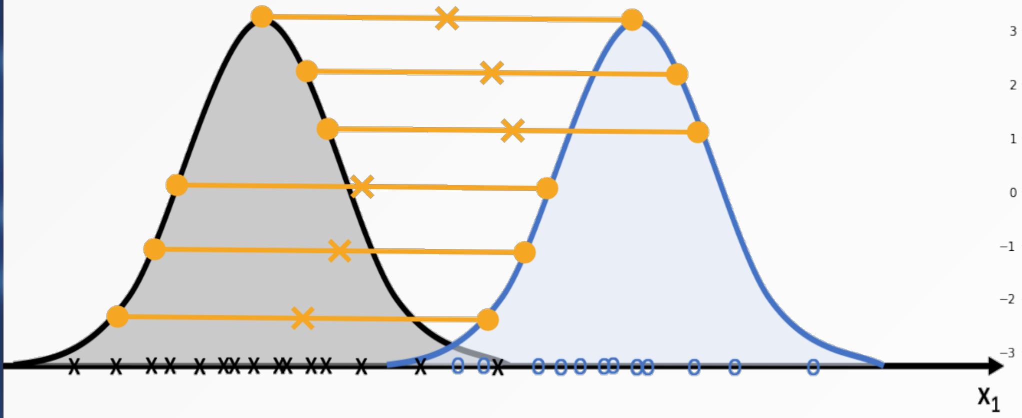


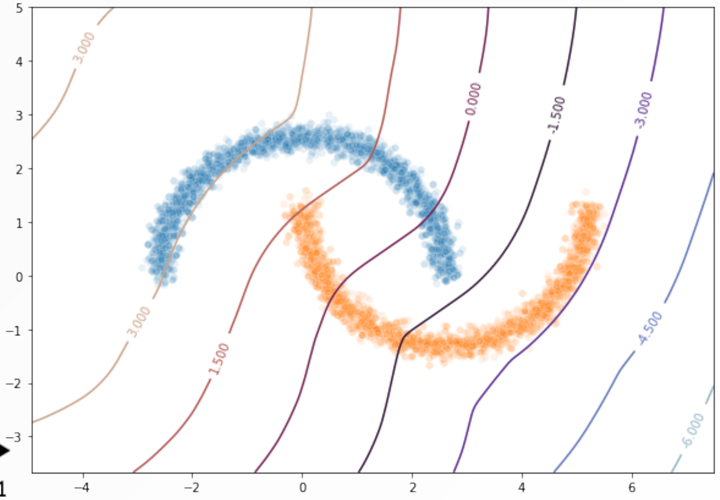
$$\mathscr{W}(\mu_0, \mu_1) = \sup_{f \in Lip_1(\Omega)} \mathbb{E}_{X \sim \mu_1}[f(X)] - \mathbb{E}_{X \sim \mu_0}[f(X)]$$

Regularization the transport cost with Hinge classification loss

The function **f** is a weak classifier for two-class classification problem : robust (1-Lipschitz) but insufficient classification performance.

$$\mathscr{W}(\mu_0, \mu_1) = \sup_{f \in Lip_1(\Omega)} \mathop{\mathbb{E}}_{x \sim \mu_1} [f(x)] - \mathop{\mathbb{E}}_{x \sim \mu_0} [f(x)]$$





Regularization the transport cost with Hinge classification loss

ANITI Université Fédérale Toulouse Midi-Pyrénées

Enforcing the classifier to

Separate the classes and

Still performing a mass

Transport between

Distributions

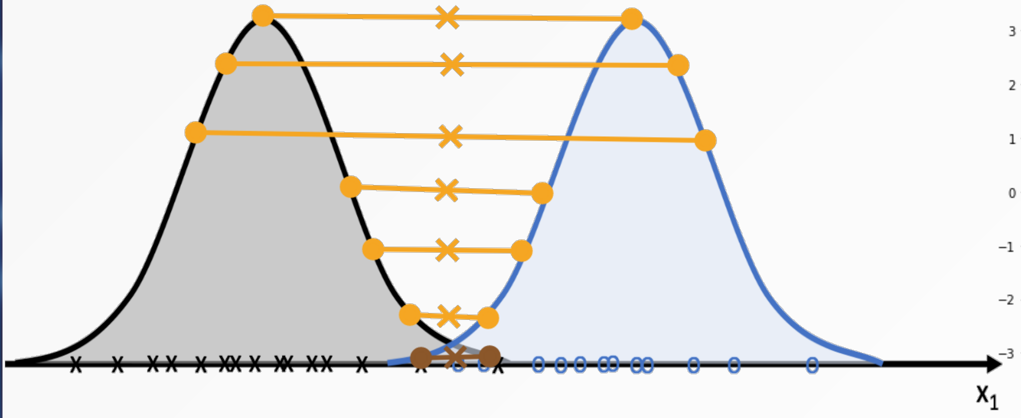
The function f is a weak classifier for two-class classification problem : robust (1-Lipschitz) but insufficient classification performance.

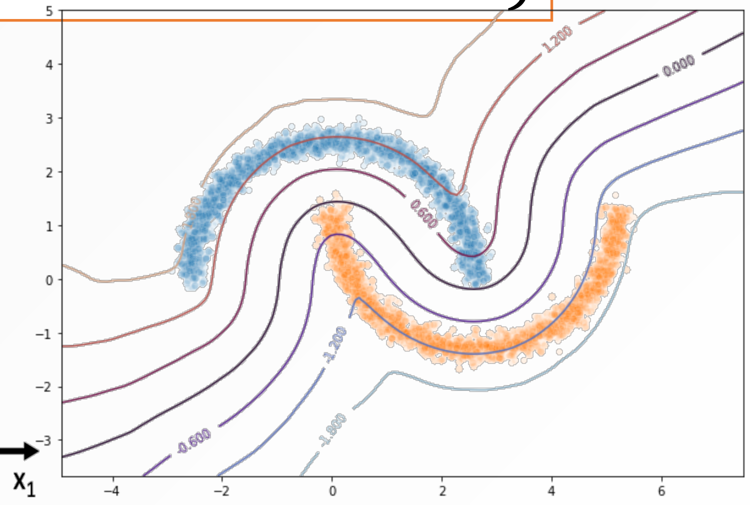
$$\mathscr{W}(\mu_0, \mu_1) = \sup_{f \in Lip_1(\Omega)} \mathop{\mathbb{E}}_{X \sim \mu_1} [f(x)] - \mathop{\mathbb{E}}_{X \sim \mu_0} [f(x)]$$

Novelty : Using Optimal Transport to classify as a natural Lipschitz classifier

Our HKR loss allow the training of robust and performant classifiers

$$\inf_{f \in Lip_1(\Omega)} \left\{ \mathop{\mathbb{E}}_{X \sim \mu_0} [f(X)] - \mathop{\mathbb{E}}_{X \sim \mu_1} [f(X)] + \lambda \mathbb{E}_P \left(1 - Y f(X)\right)_+ \right\}$$



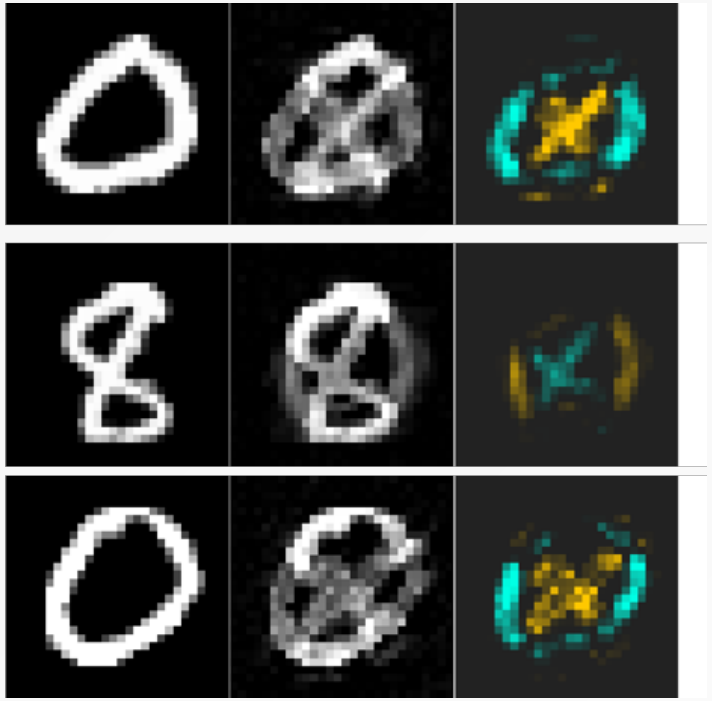


Adversarial Attacks become counterfactually reasonable examples

Theorem : the classifier can only be attacked by following the transport plan which is the direction of the gradient leading to counterfactual explanations.

$$adv(\hat{f}, x) - x = c_x \cdot \hat{f}(x) \cdot \nabla_x \hat{f}$$

Hybrid-AI : OT-based interventions with **applications to fairness** (poster N. Asher L. De Lara*)



ANITI's synergy in Deel's Project on Certifiable AI

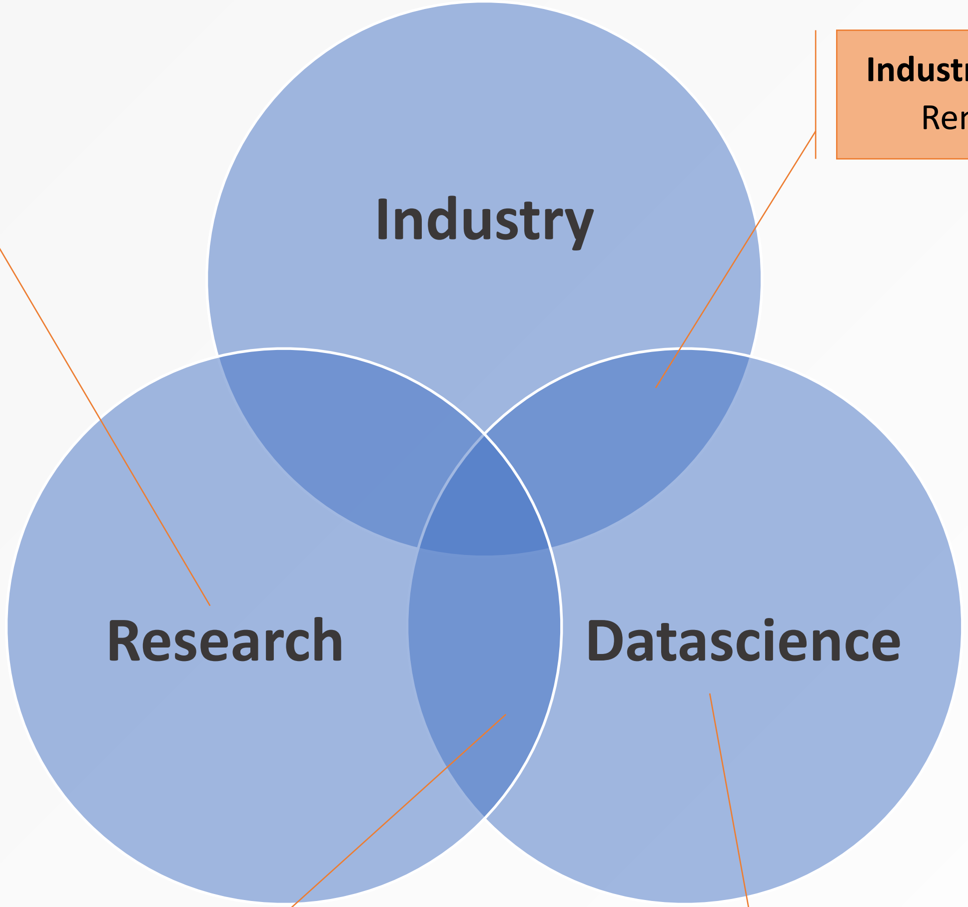
1/ Co-Designing feasible certifiable algorithms corresponding to industrial requisites

2/ Multidisciplinary Research

3/from Theory to Industrial Applications

Theory on Optimal transport
Existence and uniqueness as a transport

Industrial dataset expertise
Renault-Thales-SNCF



Theory on Machine Learning
Generalisation and rates

Practical Implementation
Python open source package

FROM RESEARCH TO INDUSTRIAL APPLICATIONS

deel-lip

DEEL LIP

LIPSCHITZ KERAS LAYERS

latest

Search docs

CONTENTS:

Example and usage

Demo 1: Wasserstein distance estimation on toy example

Demo 2: HKR Classifier on toy dataset

Demo 3: HKR classifier on MNIST dataset

⊟ deel.lip package

deel.lip.activations module

deel.lip.callbacks module

deel.lip.constraints module

deel.lip.initializers module

deel.lip.layers module

deel.lip.losses module

deel.lip.model module

deel.lip.normalizers module

deel.lip.utils module

» deel.lip package » deel.lip.layers module

Edit on GitHub

deel.lip.layers module

This module extends original keras layers, in order to add k lipschitz constraint via reparametrization. Currently, are implemented:

- **Dense layer:**

  as SpectralDense (and as FrobeniusDense when the layer has a single output)

- **Conv2D layer:**

  as SpectralConv2D (and as FrobeniusConv2D when the layer has a single output)

- **AveragePooling:**

  as ScaledAveragePooling

- **GlobalAveragePooling2D:**

  as ScaledGlobalAveragePooling2D

By default the layers are 1 Lipschitz almost everywhere, which is efficient for wasserstein distance estimation. However for other problems (such as adversarial robustness) the user may want to use layers that are at most 1 lipschitz, this can be done by setting the param *niter_bjorck=0*.

class `deel.lip.layers.Condensable`

Bases: `abc.ABC`

Some Layers don't optimize directly the kernel, this means that the kernel stored in the layer is

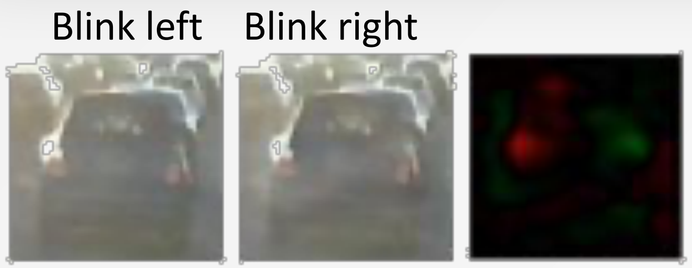
open source library

```
model = Sequential(
    [
        Input(shape=(28, 28, 1)),
        SpectralConv2D(16, (3, 3), activation=GroupSort(2), use_bias=True),
        ScaledAveragePooling2D(pool_size=(2, 2)),
        SpectralConv2D(16, (3, 3), activation=GroupSort(2), use_bias=True),
        ScaledAveragePooling2D(pool_size=(2, 2)),
        Flatten(),
        SpectralDense(32, activation=GroupSort(2), use_bias=True),
        SpectralDense(10, activation=None, use_bias=False),
    ],
    name="hkr_model",
)

model.compile(
    loss=HKR_multiclass_loss(alpha=5, min_margin=0.25),
    optimizer=Adam(lr=0.001),
    metrics=["accuracy"],
)
```

Specific training for ANITI's partners

From research to industrial applications

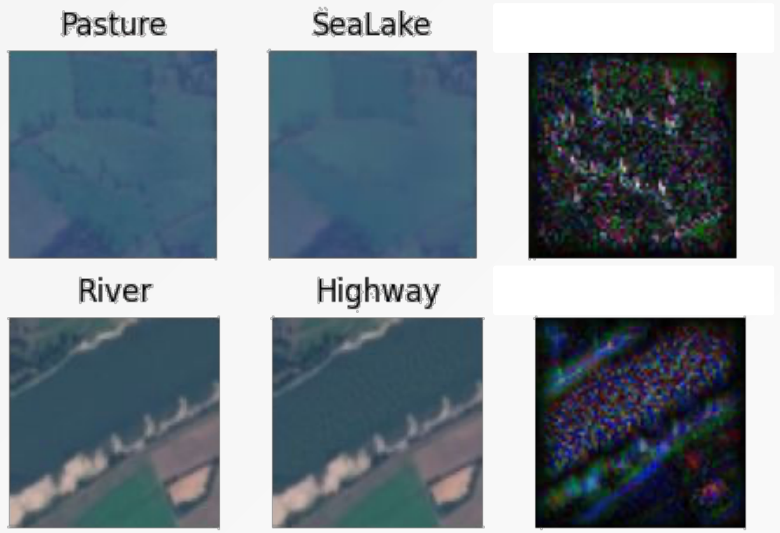


Blink classification

- Real-world dataset
- Low resolution images (32x32 RGB)
- Small dataset (frugal learning)

FRSign railway classification

- cropped images to 64x64
- binary or multiclass problem



classification of satellite image patches

- Eurosat dataset
- 64x64 image patches
- Multiclass problem (10 classes)

Thank you

For your attention