

# DEEL

DEpendable & Explainable Learning

DEpendable & Explainable Learning





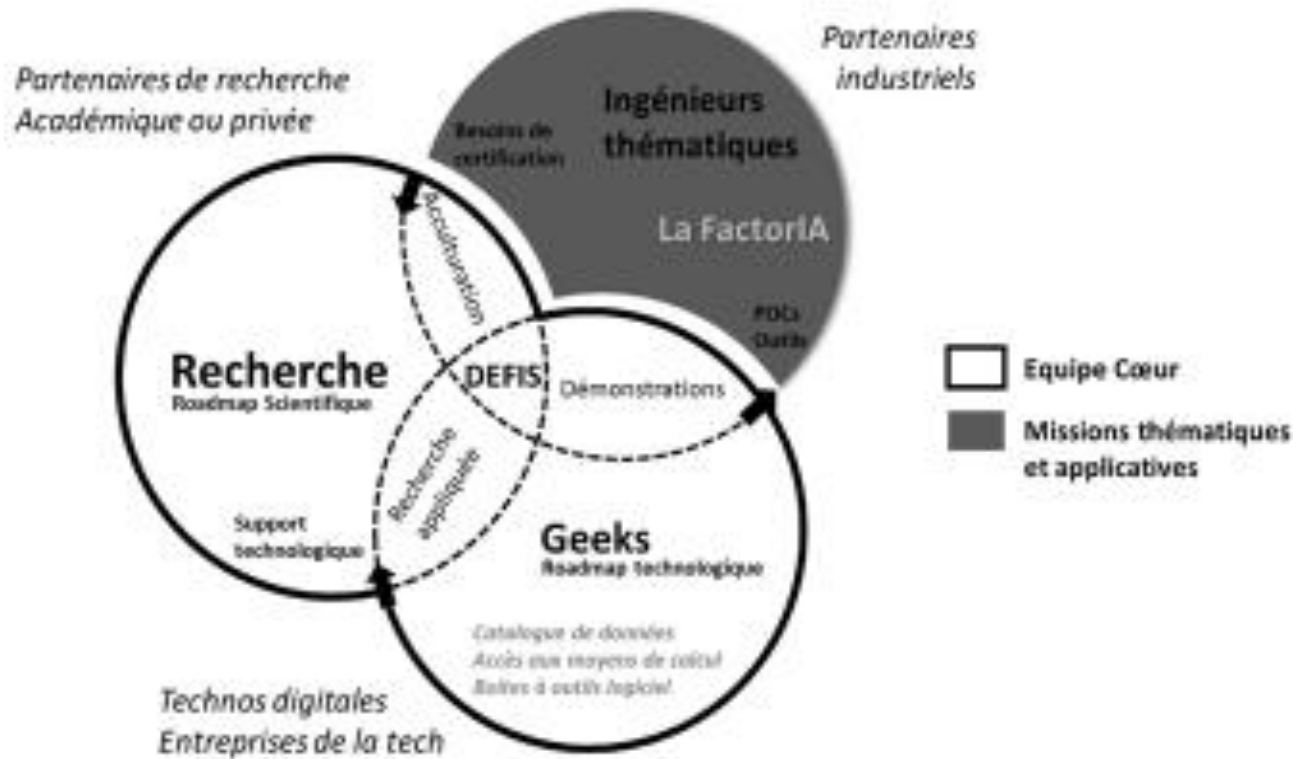
# FROM END 2018...



# ...TO END 2023

# ORGANISATION FR

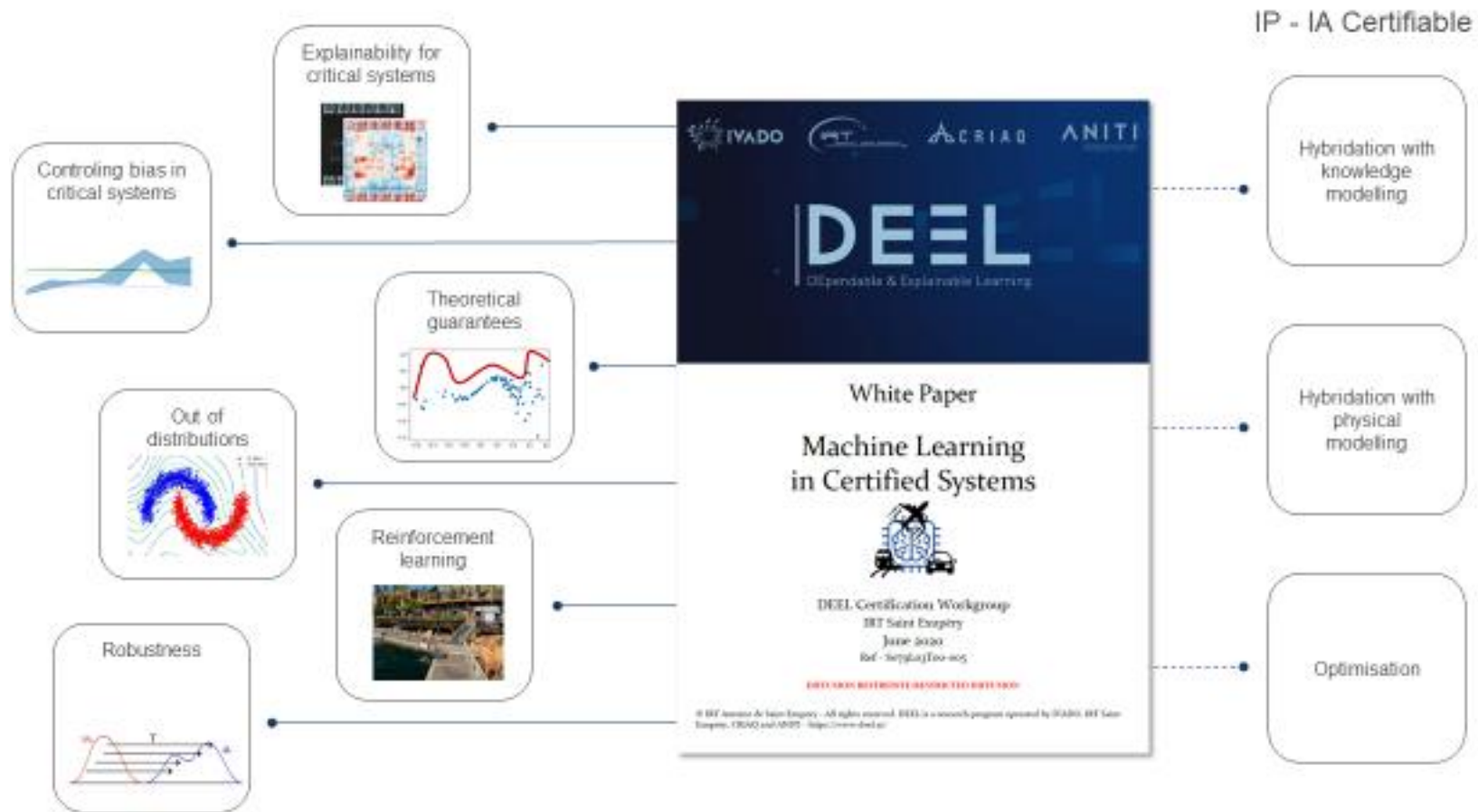
## Principe de colocalisation des équipes



Antenne à Montréal



# DEEL - CHALLENGES





# EXPLAINABILITY CHALLENGE

DEpendable & Explainable Learning



# EXPLAINABILITY CHALLENGE

- Kickoff : 2019 October 31<sup>th</sup>
- Team: 7 persons + 1 Phd + 2 researchers

# OUTPUT

- State of the art document
  - Audiences (Who ?)
  - Explainability type of output (What ?)
  - Problematics (Why ?)
  - Design process (When ?)
  - Industrial Uses-cases
  - Toolboxes
  - **Mapping between the technics and the audience, problematics and industrial usecases**
  - State of the art description
- Current Results:
  - Explainability toolbox
    - Local
    - Global
    - Metrics
  - Notebooks to evaluate technics



# WHY DO WE NEED EXPLANATIONS



# Why do we need Explanations

---

Build trust in the model prediction [3][4]



Elucidate important aspects of learned models [4]



Help satisfy regulatory requirements and Certification process[1]



Reveal bias or other unintended effects learned by a model [3]



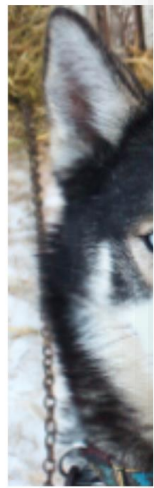
[1] Bryce Goodman & al. European union regulations on algorithmic decision-making and a "right to explanation".

[2] Finale Doshi-Velez & al. Accountability of ai under the law: The role of explanation

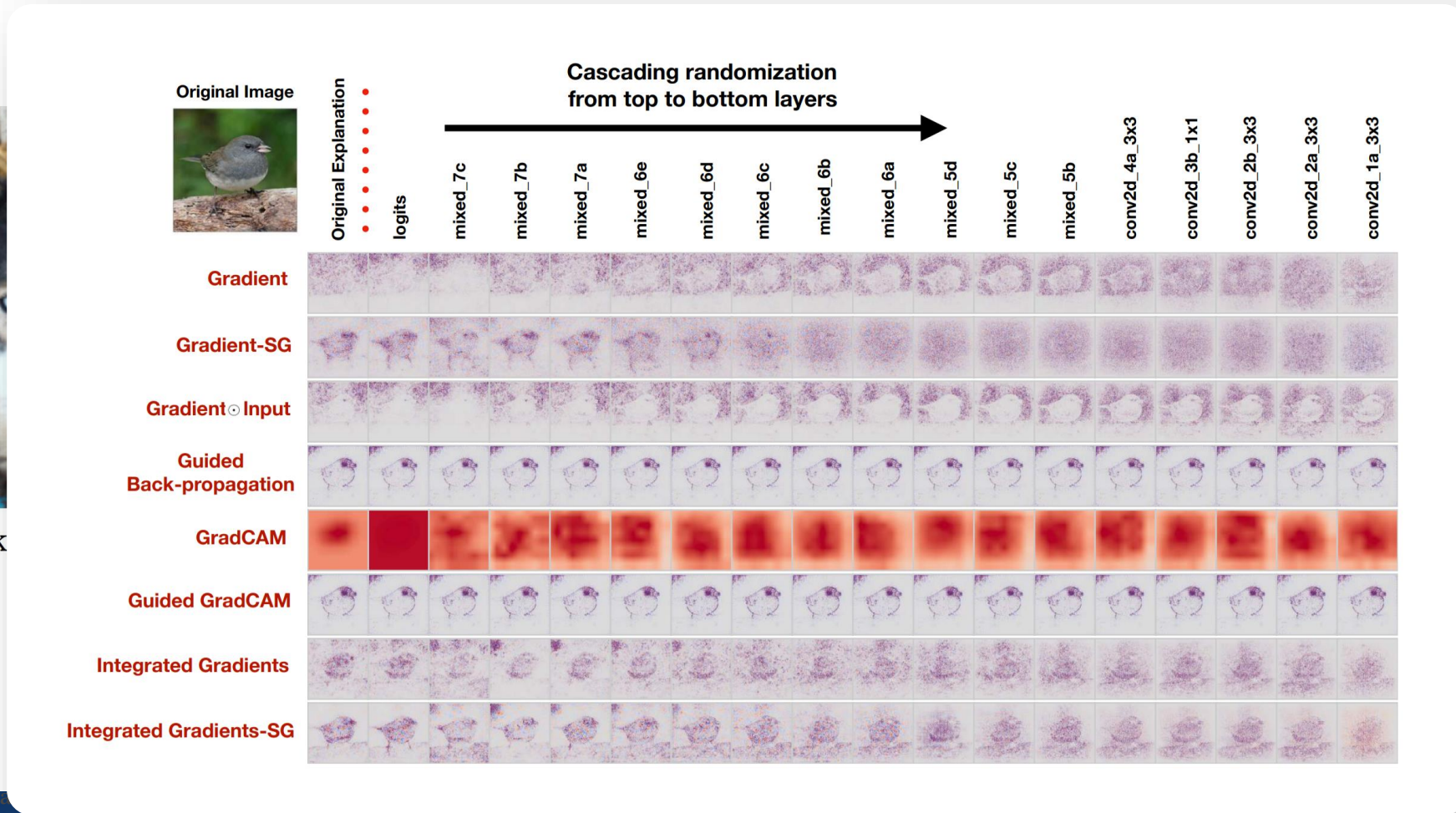
08/04/2021 [3] Gabriel Cadamuro & al. Debugging machine learning models

[4] Alfredo Vellido & al. Making machine learning models interpretable

# “What is a good explanation?”



(a) Husk



© DEEL - All rights reserved to IVADO, IRT Saint Exupéry, CRIAQ and ANITI. Confidential and proprietary document

Confirmation bias.

**Just because it makes sense to humans doesn't mean it reflects the evidence for prediction.**

# STATE OF THE ART



# STATE OF THE ART OVERVIEW (1/2)

- Global explanations
  - Transparency models
  - Features relevance explanations
  - Explanation by simplification
  - Internal analysis
  - Explanation by examples
  - Natively explainable models
    - Models providing an explanation as output
    - Building interpretable features
    - Attention models
    - Unsupervised learning for representation disentanglement

# STATE OF THE ART OVERVIEW (2/2)

- Global explanations:
  - Causality
  - Formal methods
- Local explanation
- Validation:
  - Metrics
  - Explainability Robustness
  - Link between Robustness and Explicability

# WORKING AXES





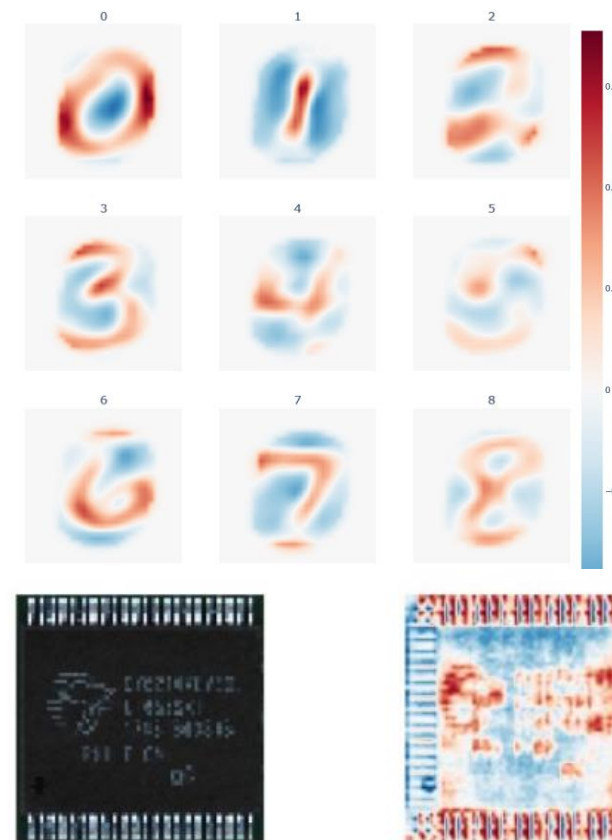
## 2 MAIN AXES

- Research thematics:
  - Goal: Develop research axes which are important for Deel project and which are not much investigated in the research community
- Deel Explainability “Library”: Evaluation of existing technologies on our industrial usecases
  - Goal: Create software suite and Jupyter notebook tutorials
    - Tutorials are given to explain how Explainability techniques shall be used to analyse different industrial uses cases
    - The techniques could be implemented in a DEEL library or relying on existing external toolboxes

# RESEARCH THEMATIC

- Outcomes : Articles & source code
- **Metrics / Explainability Robustness:**
  - 2 core team members
  - 1 researcher
  - 1 PhD student
- **Formal methods**
  - 2 core team members
  - Link to ANITI project
- Backlog:
  - Causality
  - Link between Robutness and Explainability
  - Building interpretable features & Attention models

Library for Global Explanation:  
[www.gems-ai.com](http://www.gems-ai.com)

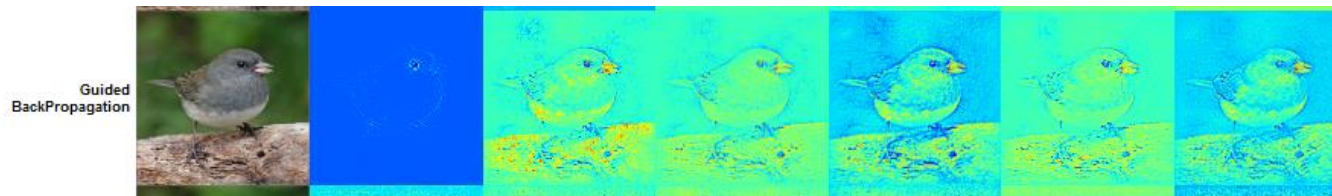


# DEEL EXPLAINABILITY “LIBRARY”

- Outcome:

- Deel Explainability Library (source code)
- Tutorials (Jupyter notebook)
- Feed back on industrial usecases

## Internal model analysis



1	Internal Analysis
2	Building features/ attentions / Unsupervised learning for representation disentanglement
3	Formal methods
4	Inputs local Importance
5	Causality
6	Link between explicability / Robustness
7	Metrics



First evaluation done



3 core team members



2 core team members

## VAE Evaluation on Deel Dataset

Amount of pins



90 degree rotation



# REPRESENTATIVITY AND CONSISTENCY MEASURES FOR DEEP NEURAL NETWORK EXPLANATIONS



## Properties of explainability

### Fidelity

Does my explanation reflect the behavior of my model?

### Representativity

How many phenomena my explanation cover?

### Comprehensibility

Is my explanation unambiguous and simple?

### Consistency

The degree to which similar explanations are generated from different models trained on the same task.

### Stability

Does my explanation remain the same under semantically invariant transformation?

### Novelty

Does my explanation reflect the fact that explained instance is from a new region, not contained or well represented in the training set?

### Consistency

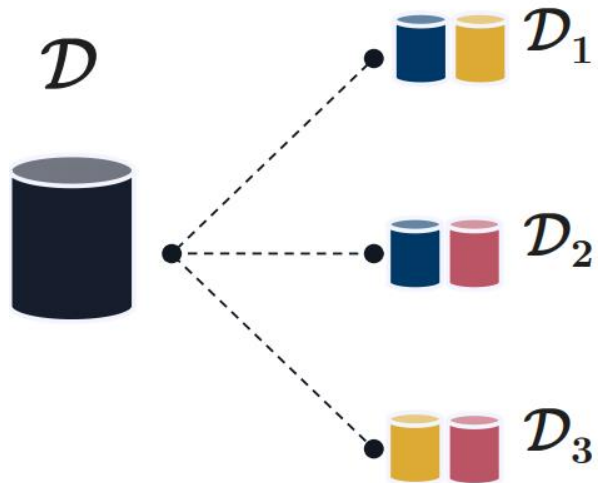
An explanation leading to predict  $y$  and  $\neg y$  is inconsistent.

### Representativity

A model should not base an explanation on a single sample.

# TRAINING PROCEDURE

Split the dataset



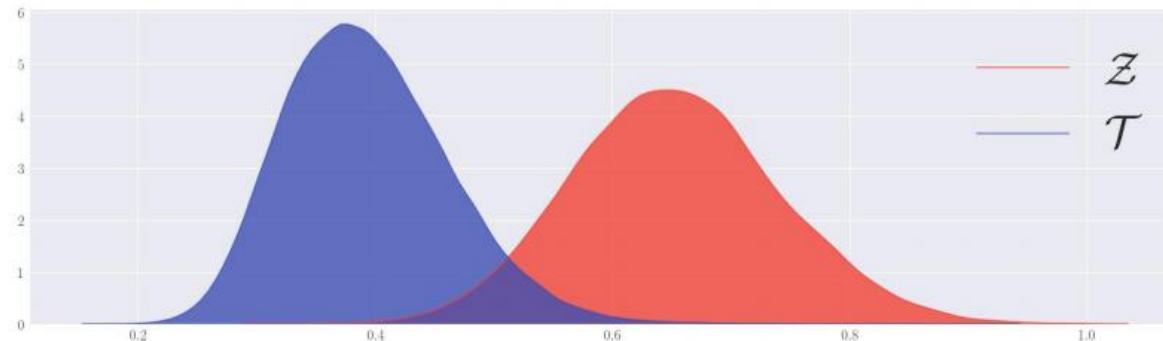
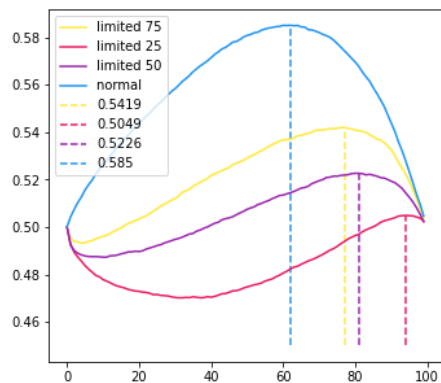
$$\forall (x, y) \in \mathcal{D}, \forall \mathcal{D}_i : x \in \mathcal{D}_i, \forall \mathcal{D}_j : x \notin \mathcal{D}_j$$

$$\mathcal{T} = \{d(\phi_x^{\mathcal{D}_i}, \phi_x^{\mathcal{D}_j}) \mid f_{\mathcal{D}_i}(x) = y \wedge f_{\mathcal{D}_j}(x) = y\}$$

$$\mathcal{Z} = \{d(\phi_x^{\mathcal{D}_i}, \phi_x^{\mathcal{D}_j}) \mid f_{\mathcal{D}_i}(x) = y \oplus f_{\mathcal{D}_j}(x) = y\}$$

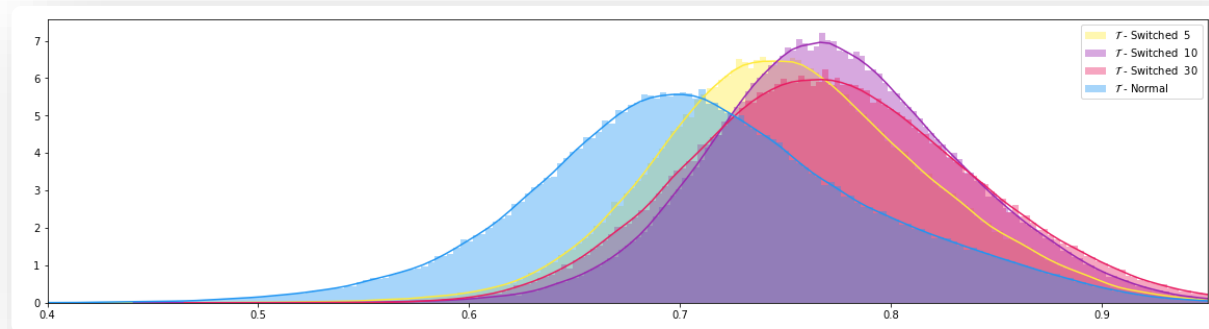
# RELATIVE CONSISTENCY

$$ReCo = \max_{\gamma} TPR(\gamma) + TNR(\gamma) - 1$$



# MEAN GENERALIZABILITY

$$MeGe = \frac{1}{1 + \mathbb{E}[\mathcal{T}]}$$

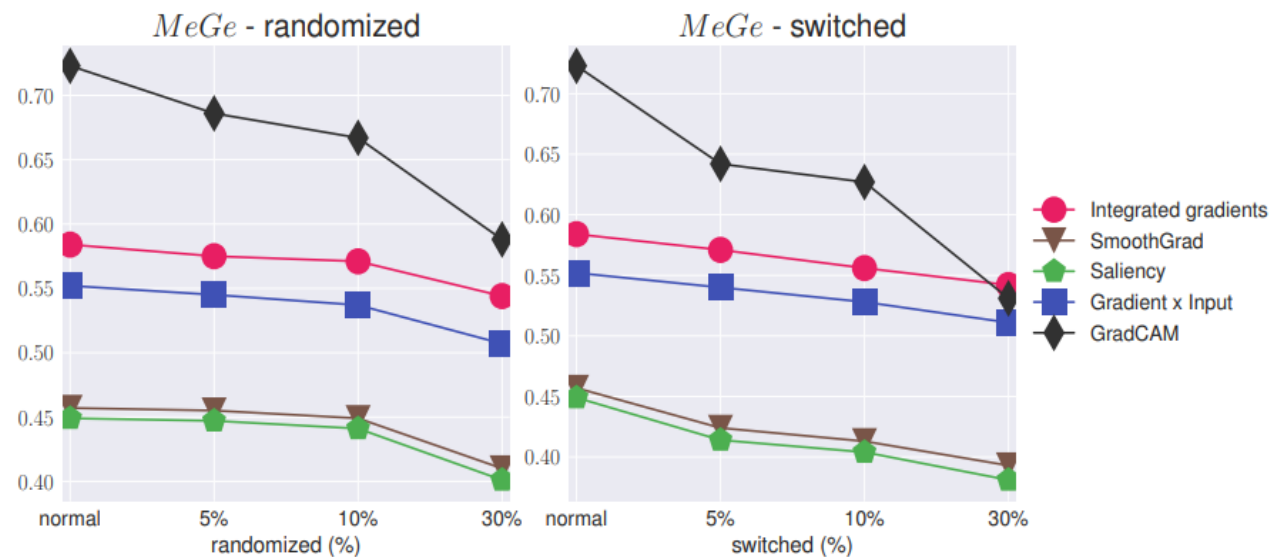




# EVALUATE METRIC IN FRONT OF MODELS DEGRADATIONS

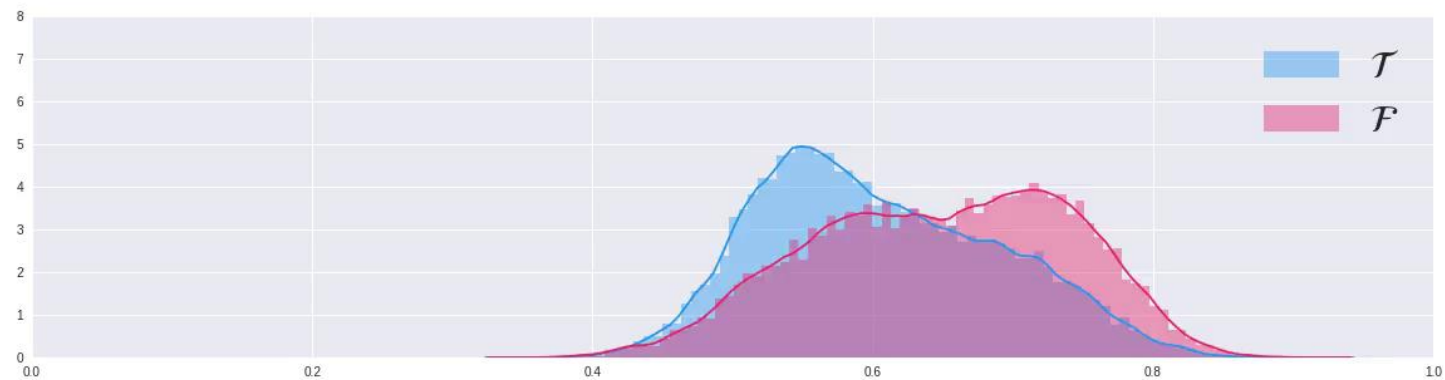
ReCo scores for normally trained ResNet-18

Model - Dataset	IG	SG	SA	GI	GC
Cifar10	0.107	0.154	0.151	0.088	<b>0.637</b>
Cifar100	0.018	0.132	0.131	0.004	<b>0.800</b>
EuroSAT	0.309	0.182	0.177	0.241	<b>0.591</b>
FashionMNIST	0.369	0.125	0.1	0.369	<b>0.517</b>



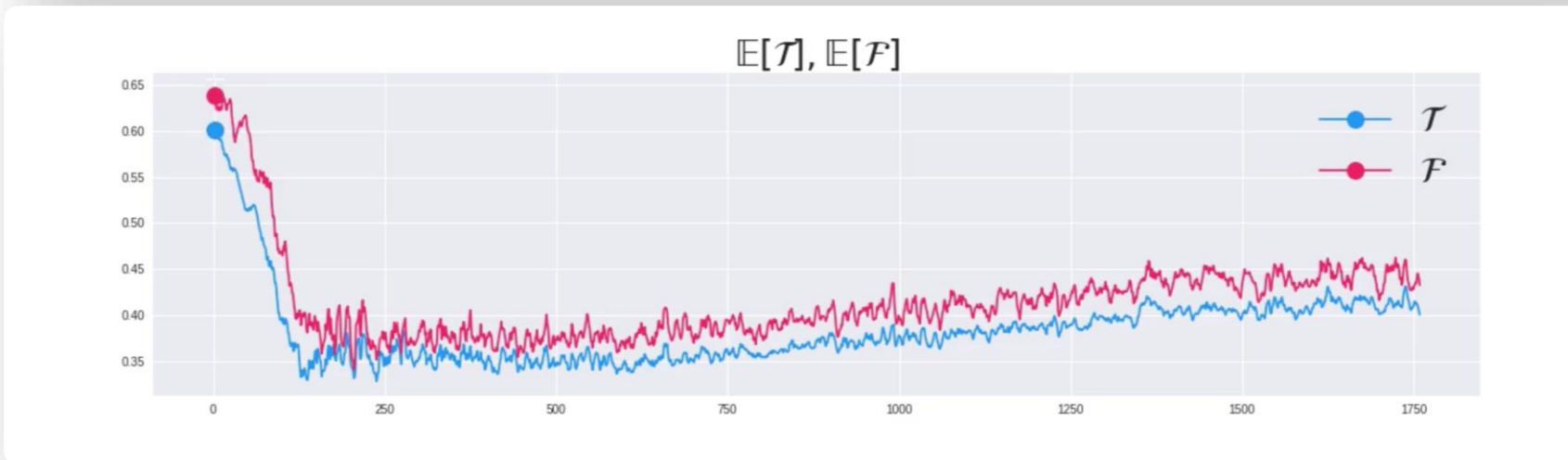
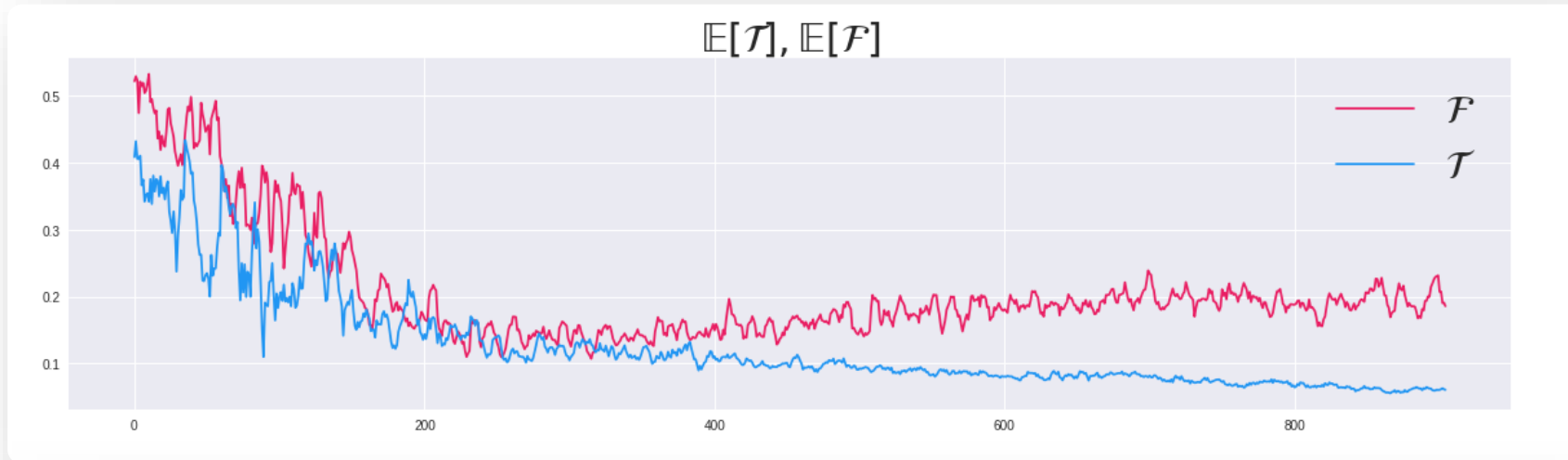
# OBSERVATIONS

Fashion MNIST



# OBSERVATIONS

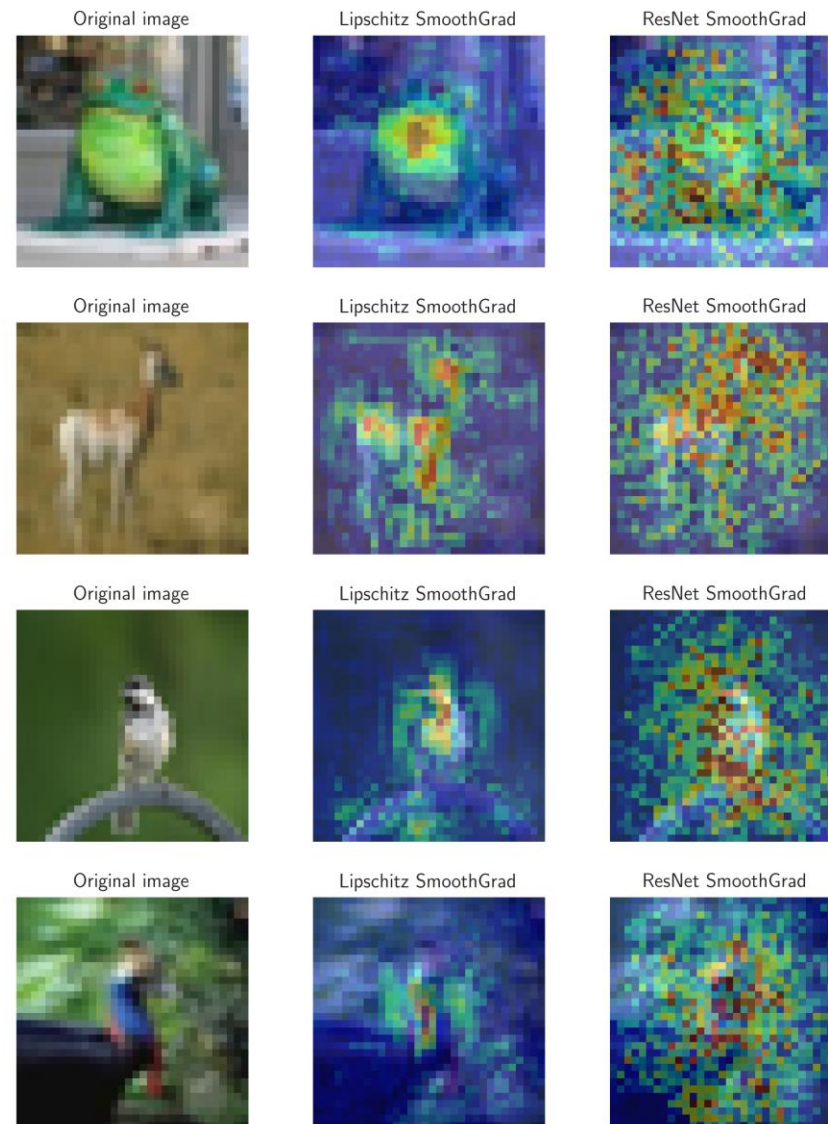
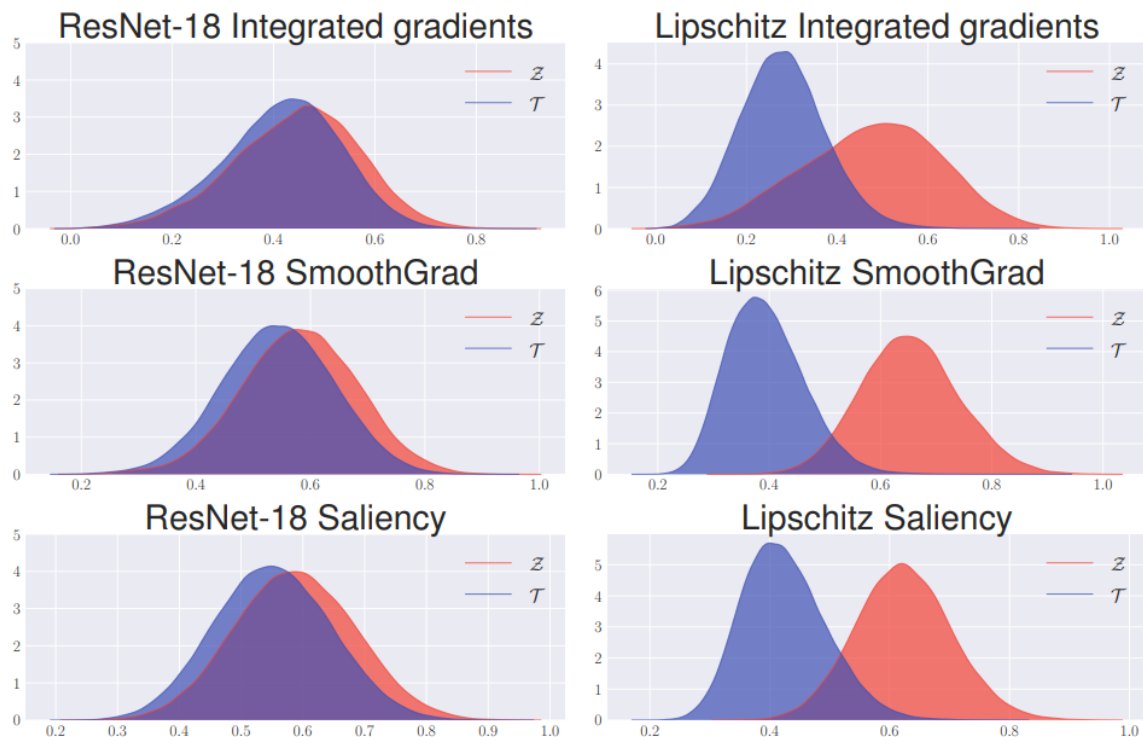
Fashion MNIST



# LIPSCHITZ

	IG	SG	SA	GI	GC	Shap
<b>Lipschitz</b>	<b>0.598</b>	<b>0.898</b>	<b>0.81</b>	<b>0.5</b>	<b>0.668</b>	0.38
ResNet-18	0.107	0.154	0.151	0.088	0.637	<b>0.387</b>

Table 2. *ReCo* score obtained on Cifar10. Higher is better.



# Preprint: <https://arxiv.org/abs/2009.04521>

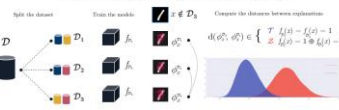


Figure 1: Example of application of the proposed procedure, with  $D_1$  dataset and the number of splits  $S = 3$ , each of the models is trained on two of the data splits. For a given sample  $x$  such that  $x \in D_1$ , the explanations for each model are calculated ( $e_1^x, e_2^x, e_3^x$ ), and the distances between the explanation of the model not trained on  $x$  ( $d_{21}^x, d_{31}^x$ ) and the two models trained on  $x$  ( $d_{12}^x, d_{13}^x$ ) are computed. For each distance, if the prediction of both models are correct, the distance is added to  $T$ , if one of the two models makes a false prediction, the distance is added to  $Z$ .

prediction, if the distance between two explanations is small it means that the model built its explanations from several samples, and that the fact of having or removing a particular object does not make these explanations vary, which is a sign of good generalization. In the case where one of the two models gives a bad prediction, we want to avoid that they give the same explanation, indeed the consistency of the explanations means that we cannot justify with the same explanation two different outcomes. Hence, we observe two cases, where the models were right where we used a small distance between the explanations, and where one of the models gives a wrong prediction where we used a high distance between the two explanations. We build two

of associated metrics such as consistency or generalization. This work proposes a methodology applicable to a large family of models, based on a distance between explanation coming from a same sample. We use this methodology to introduce two new metrics, Relative Consistency ( $RC_C$ ), motivated by the idea that an explanation should be used to justify non-contradictory decisions, and use it to measure the consistency of the explanations. The second metric, Mean Generalizability ( $MG_G$ ), is intended to measure the ability of a model to derive general rules from its explanations.

We then use the new procedure into practice in order to calculate the two metrics on different datasets. The results obtained between normally trained models and degraded models allow us to assess that the metrics reflect the loss of generalization and consistency of the explanations. Finally, we use these metrics to highlight in a quantitative way the suggestions of different models on the generalization and coherence capacity of Laplace networks.

- 2. Related Works**
- 2.1. Related Works**

Recently, several works have proposed methods to explain the results of DNN, i.e. to make them interpretable and understandable to humans. These methods use a wide variety of tools specific to each domain. They can nevertheless be



Figure 2: ClfRR  $MG_G$  scores for normally trained and degraded models. On the left, as well as for progressively degraded models. On the left, the results are progressively more degraded, on the right the results are trained with inverted labels.

models (Eshoh et al., 2017; Adhikary et al., 2018; Toussaint et al., 2018). However, it is important to note that the visual similarity problem is still an open problem. This is why, in order to corroborate the results obtained, we tested several distances, among which: Wasserstein, Soft-Dice, Spearman rank,  $SSIM$ ,  $L_1$  and  $L_2$  norms, all giving consistent results between them. We also conduct various sanity checks to ensure they could respond to the problem (see additional materials A.1). Considering the computation time, the numerical stability of the distance as well as taking into account the results of the tests carried out, we finally chosen to stay in line with previous work by essentially using the absolute value of Spearman rank correlation.

**5. Results**

We evaluated the  $RC_C$  and  $MG_G$  metrics using the procedure introduced in 3.1, with the parameters and constraints set in 4.1 on the various datasets described in 1.1. For each of these datasets, the  $RC_C$  and  $MG_G$  scores were calculated for normally trained models, as well as for models having degradations at three different degrees (described in 2.1). In addition, on the ClfRR dataset we extended the list of methods by adding SHAP analysis (Lice, 2017) and trained a family of Laplace models known for their good generalization capability (Vehring et al., 2015; Turner et al., 2010) using DEEL library.

The purpose of this section is to report and synthesize the main results for a complete detail, refer to the appendix A.1, which are essentially three observations. The first concerns the variation in the  $MG_G$  and  $RC_C$  scores according to the different explanation methods used, the second observation concerns the differences in  $MG_G$  scores between the normally trained and degraded models, and finally the importance of the model model couple, where we observe that certain methods are better suited to a family of models with a tendency regarding the Grad-CAM method that seems robust to different families.

**5.1. Methods ranking**

Model	Dataset	IG	SH	SA	CI	OC
ClfRR	10	0.107	0.114	0.111	0.091	0.437
		0.094	0.112	0.113	0.091	0.406
Hardest	100	0.309	0.072	0.177	0.221	0.891
		0.309	0.125	0.1	0.309	0.437

**5.2. Couple method**

Method	IG	SH	SA	CI	OC	Shap
Laplace	0.298	0.408	0.411	0.425	0.448	0.101
	0.268	0.407	0.416	0.421	0.447	0.092

**5.3. ClfRR model method**

Model	Dataset	IG	SH	SA	CI	OC
ClfRR	10	0.107	0.114	0.111	0.091	0.437
		0.094	0.112	0.113	0.091	0.406

**5.4. Couple method**

Method	IG	SH	SA	CI	OC	Shap
Laplace	0.298	0.408	0.411	0.425	0.448	0.101
	0.268	0.407	0.416	0.421	0.447	0.092

**5.5. ClfRR model method**

Model	Dataset	IG	SH	SA	CI	OC
ClfRR	10	0.107	0.114	0.111	0.091	0.437
		0.094	0.112	0.113	0.091	0.406

## Generalization & Consistency Metrics from Deep Neural Network Explanations

### Metric term 1<sup>2</sup>

However, their advantages come with major drawbacks, especially concerning the difficulty of interpreting their decisions, and they are considered as black box models (Lipton, 2016). This problem is a very serious obstacle to the adoption of these systems in a so-called critical context such as aeronautics or medicine, and it is widely recognized that this adoption will not occur without proper explanations of how these models take decisions.

Recently, many strategies have been proposed to help users understand the underlying logics that led these models to a particular decision. Some work (Adhikary et al., 2018; Ghoshbani et al., 2017) has shown the potential pitfalls associated with current interpretation methods. While some methods offer explanations that are satisfactory to users, they do not, however, reflect the real functioning of the model. The interpretability that was intended to provide confidence, is itself questionable.

These observations have given rise to a need for an objective assessment of the explanations produced by these methods, thus enabling benchmarks and points of reference to be established. To do this, one approach advocated by various works is to ensure that the explanations satisfy a certain number of properties (or axioms), such as fidelity, stability, generalization or consistency. Since some works propose an exhaustive list of these properties, a good explanation should be defined as quantitatively satisfactory according to a coherent set of measurements specific to each of these properties.

The machine learning community has responded to this need by proposing several metrics according to some of these properties. Most of these metrics consist of removing critical features from the input and measuring the prediction gap of the classifier, and some variants such where the models need to be re-trained to ensure this gap does not come from a distribution shift (Rieger & Hansen, 2020; Hoeker et al., 2018). Another strategy is to evaluate the sensitivity of the explanation to the vicinity of a point of interest, or by making sure that the explanation remains the same under semantically invariant transformations (Ancora et al., 2017).

Although these two strategies are promising approaches for quantifying the fidelity and stability of an explanation, there are still number of desirable properties that currently lack

**1.** By example with on the left, a model family giving good explanations, and on the right a model family with bad explanations and yet the same  $R_1$  distance.

**2.** By example with on the left, a model family giving good explanations, and on the right a model family with bad explanations and yet the same  $R_1$  distance.

**3.** By example with on the left, a model family giving good explanations, and on the right a model family with bad explanations and yet the same  $R_1$  distance.

Figure 3: Distances with moving interval point. Distances between the original image and the noisy copies computed for each value of  $\sigma$ .

Figure 4: Distances with moving interval point. Distances between the original image and the noisy copies computed for each value of  $\sigma$ .

Figure 5: Example of results obtained using the Smooth method. Both models had a correct prediction, the middle is the original image, the left and right are noisy images. The different distances evaluated pass this sanity check, i.e. the 1-Wasserstein metric ( $W_1$ )

under the cover of the  $R_1$  distance.

**A.3. Models**

As mentioned in the description of Laplace network, its size and number of filters is not fixed, but can be defined for each of filters for the  $l$  layer. The number of filters for the  $l$  layer depends on the number of filters in the previous layer. The number of filters in the  $l$  layer is denoted by  $n_l$ .

**A.4. Metrics**

The  $RC_C$  metric is defined as follows:

$$RC_C(T_1, T_2) = \log \frac{T_1 + T_2 + \epsilon}{2T_1} + \log \frac{T_1 + T_2 + \epsilon}{2T_2} - 1$$

The  $MG_G$  metric is defined as follows:

$$MG_G(T_1, T_2) = \log \frac{T_1 + T_2 + \epsilon}{2T_1} + \log \frac{T_1 + T_2 + \epsilon}{2T_2} - 1$$

The  $W_1$  metric is defined as follows:

$$W_1(T_1, T_2) = |T_1 - T_2|$$

The  $W_2$  metric is defined as follows:

$$W_2(T_1, T_2) = \sqrt{|T_1 - T_2|}$$

The  $SSIM$  metric is defined as follows:

$$SSIM(T_1, T_2) = \frac{2\mu_1\mu_2 + \epsilon}{\mu_1^2 + \mu_2^2 + \epsilon}$$

The  $SP$  metric is defined as follows:

$$SP(T_1, T_2) = \frac{|T_1 - T_2|}{T_1 + T_2 + \epsilon}$$

The  $SA$  metric is defined as follows:

$$SA(T_1, T_2) = \frac{|T_1 - T_2|}{T_1 + T_2 + \epsilon}$$

The  $CI$  metric is defined as follows:

$$CI(T_1, T_2) = \frac{|T_1 - T_2|}{T_1 + T_2 + \epsilon}$$

The  $OC$  metric is defined as follows:

$$OC(T_1, T_2) = \frac{|T_1 - T_2|}{T_1 + T_2 + \epsilon}$$

MERCI POUR VOTRE ATTENTION

