DEEL

DEpendable & Explainable Learning

IVADO — Institut de valorisation des données

IRT Saint Exupery

CRIAQ — Consortium de recherche et d'innovation en aérospatiale au Québec

ANITI — Artificial & Natural Intelligence Toulouse Institute

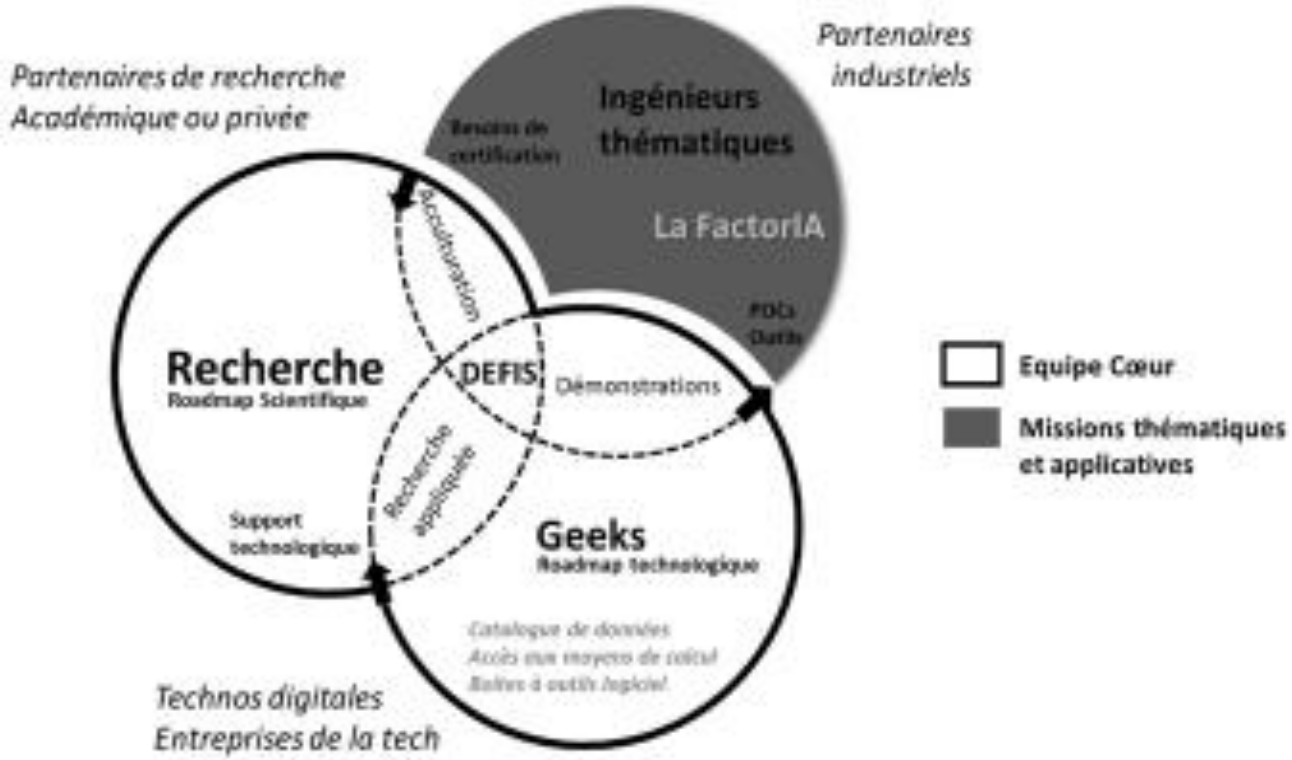
DEEL

DEEL
DEpendable & Explainable Learning

FROM END 2018…

…TO END 2023

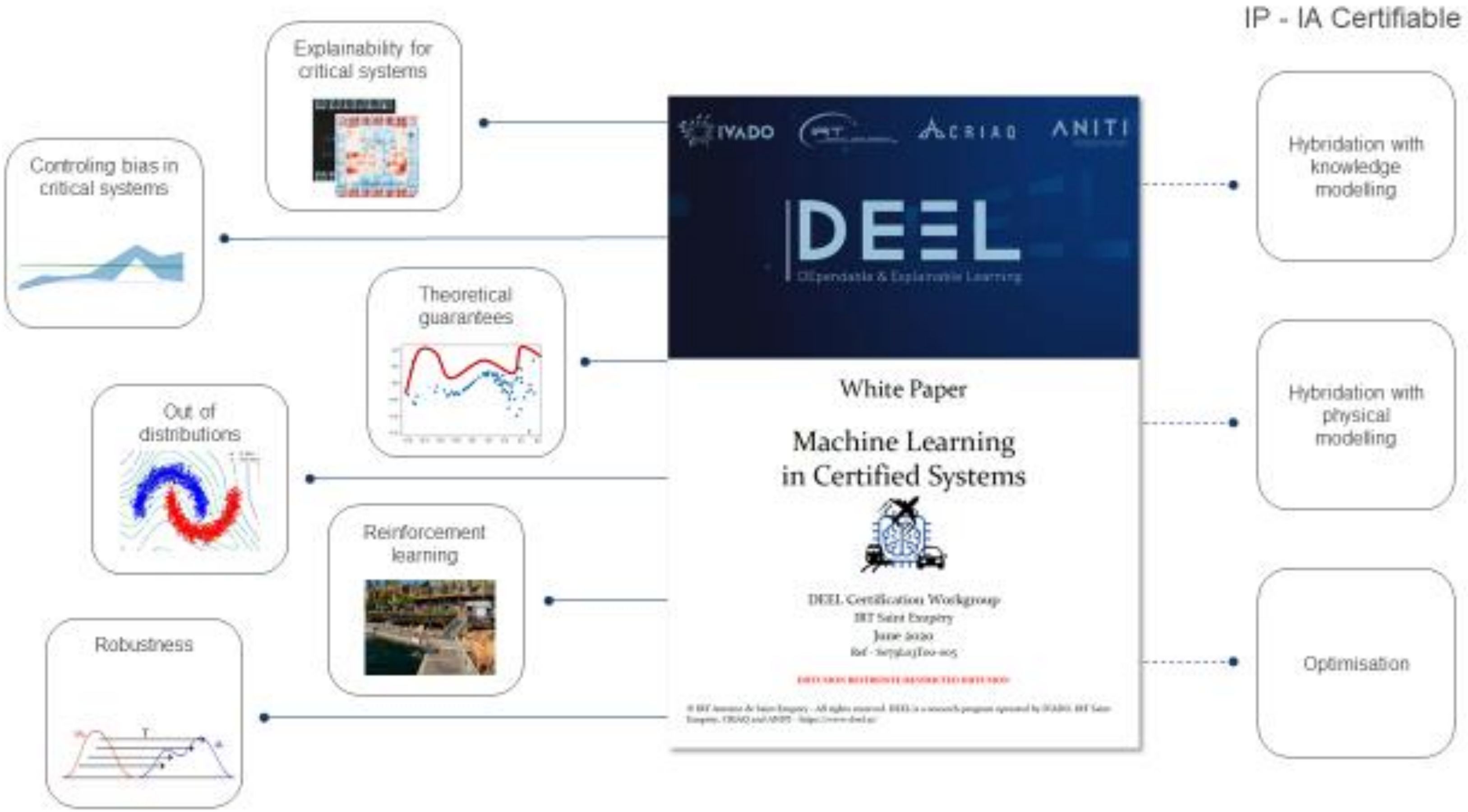
ORGANISATION FR



Principe de colocalisation des équipes

Antenne à Montréal

DEEL - CHALLENGES



DEEL

DEpendable & Explainable Learning

EXPLAINABILITY CHALLENGE

IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

EXPLAINABILITY CHALLENGE

DEEL
DEpendable & Explainable Learning

- Kickoff : 2019 October 31th
- 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

DEEL
DEpendable & Explainable Learning

WHY DO WE NEED EXPLANATIONS

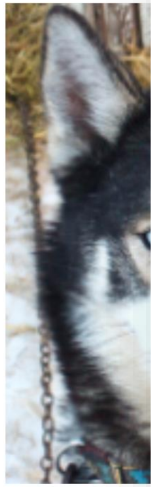
IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

Why do we need Explanations

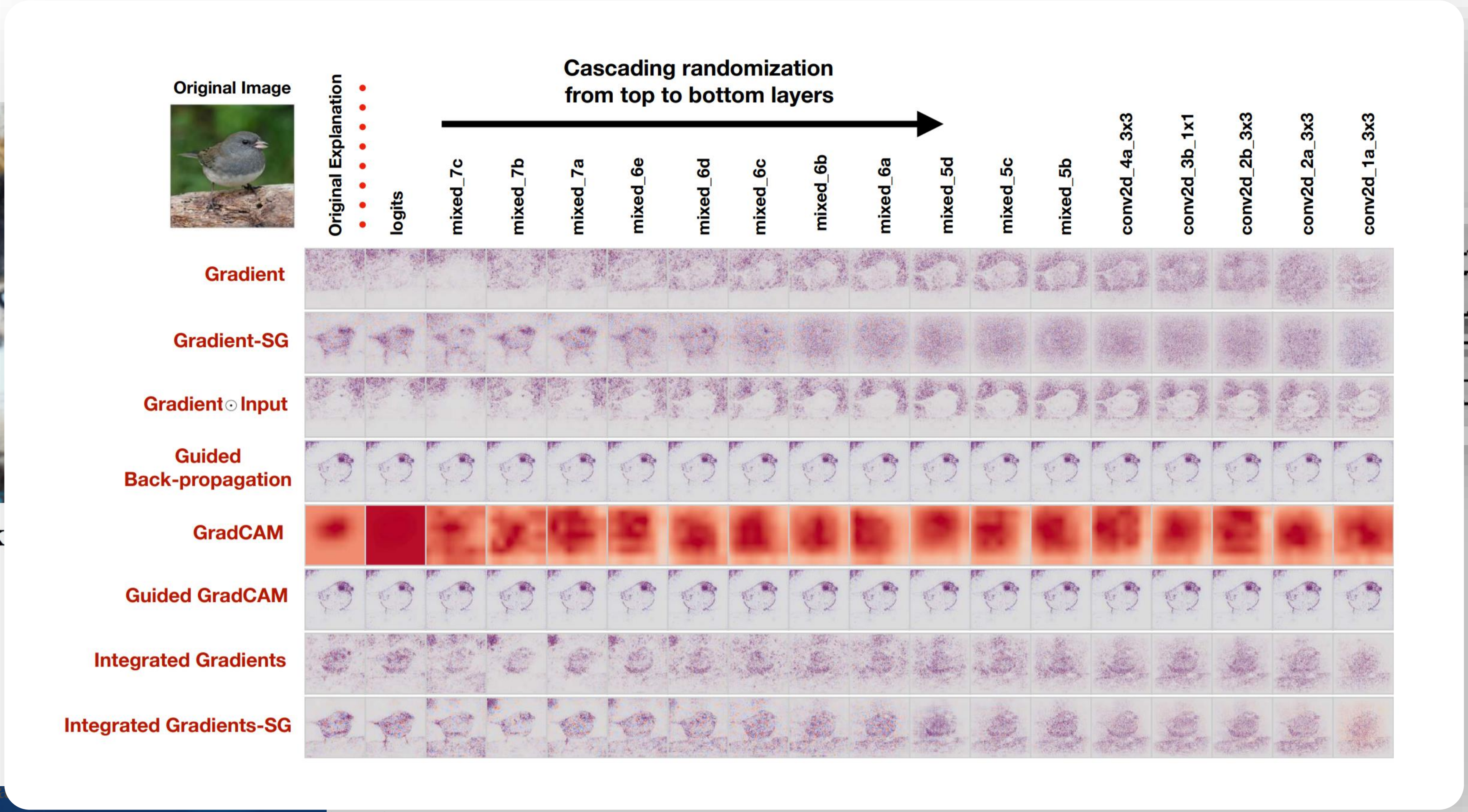
- Build trust in the model prediction [3][4]
- Elucidate important aspects of learned models [4]
- Help satisfy regulatory requirements and Certification process[1]
- Reveal bias or other unintended effects learned by a model [3]

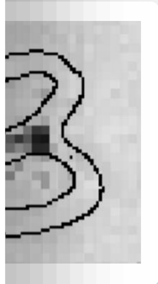
[1] Bryce Goodman & al. European union regulations on algorithmic decision-making and a"right to explanation".
[2] Finale Doshi-Velez & al. Accountability of ai under the law: The role of explanation
[3] Gabriel Cadamuro & al. Debugging machine learning models
[4] Alfredo Vellido & al. Making machine learning models interpretable

“What is a good explanation?”



(a) Husk





Explaining the Predictions of Any Classifier (2016)
Evaluating the visualization of what a Deep Neural Network has learned (2016)

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)

DEpendable & Explainable Learning

- 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

DEEL
DEpendable & Explainable Learning

WORKING AXES

IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

2 MAIN AXES

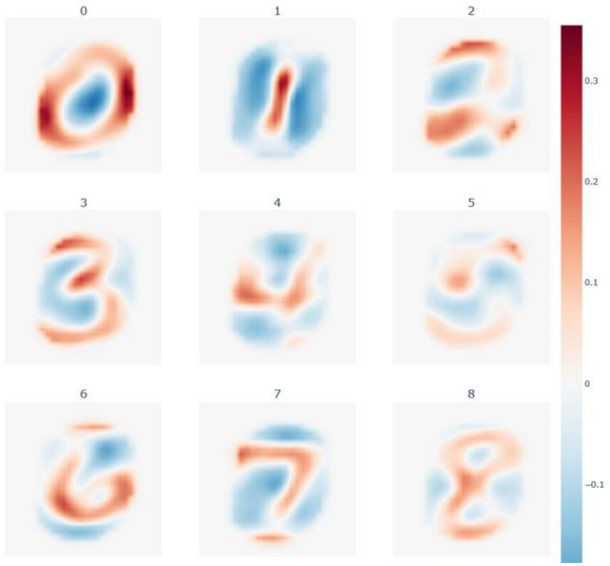
- Research thematics:
  - Goal: Develop research axes which are important for Deel project and which are not much investigated in the research community

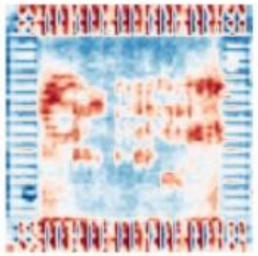
- Deel Explainability "Library": Evaluation of existing technologies on our industrial usecases
  - Goal: Create software suite and Jupyter notebook tutorials
    - Tutorials are given to explain how Explainability techniques shall be used to analyse different industrial uses cases
    - The techniques could be implemented in a DEEL library or relying on existing external toolboxes

RESEARCH THEMATICS

- 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 Robustness and Explainability
    - Building interpretable features & Attention models

Library for Global Explaination:
www.gems-ai.com





DEEL

DEpendable & Explainable Learning

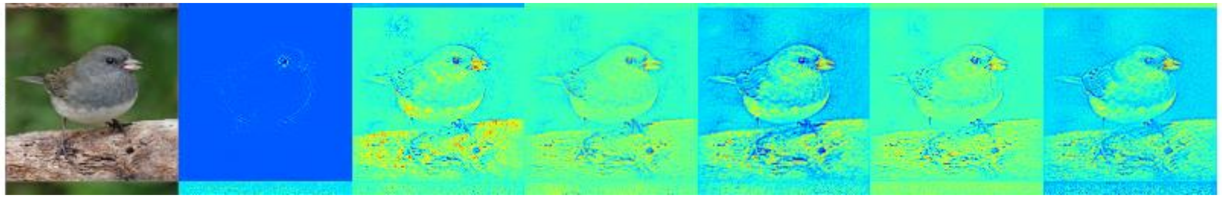
DEEL EXPLAINABILITY “LIBRARY”

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

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

Internal model analysis

Guided BackPropagation



First evaluation done

3 core team members

2 core team members

VAE Evaluation on Deel Dataset

Amount of pins

90 degree rotation

DEEL
DEpendable & Explainable Learning

REPRESENTATIVITY AND CONSISTENCY MEASURES FOR DEEP NEURAL NETWORK EXPLANATIONS

IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

Properties of explainability

DEEL
DEpendable & Explainable Learning

no metrics associated

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?

Explanation in Artificial Intelligence: Insights from the social sciences **(2019)**
Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges **(2018)**
Evaluating Explanation Without Ground Truth in Interpretable Machine Learning **(2019)**
Perturbation-Based Explanations of Prediction Models **(2018)**

Consistency

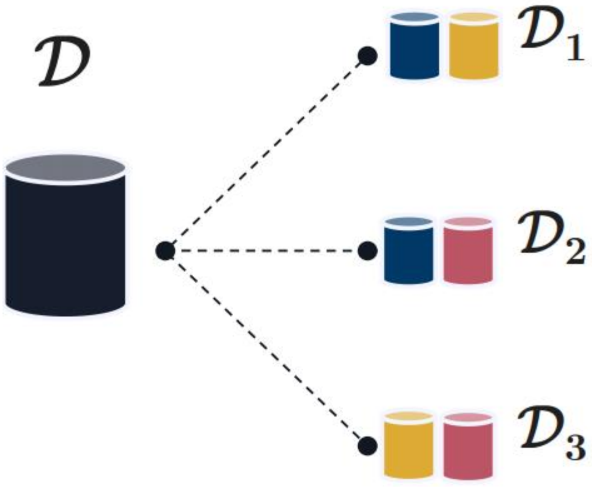
An explanation leading to predict $\mathbf{y}$ and $\neg\mathbf{y}$ is inconsistent.

Representativity

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

TRAINING PROCEDURE

Split the dataset



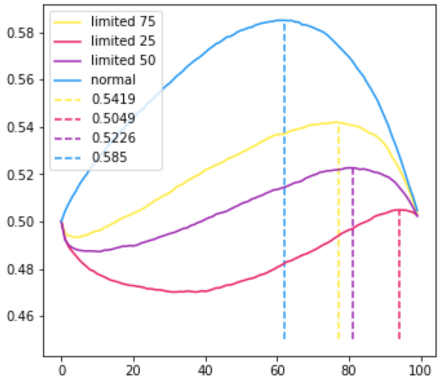
$$\forall (x,y) \in \mathcal{D},\ \forall \mathcal{D}_i : x \in \mathcal{D}_i,\ \forall \mathcal{D}_j : x \notin \mathcal{D}_j$$

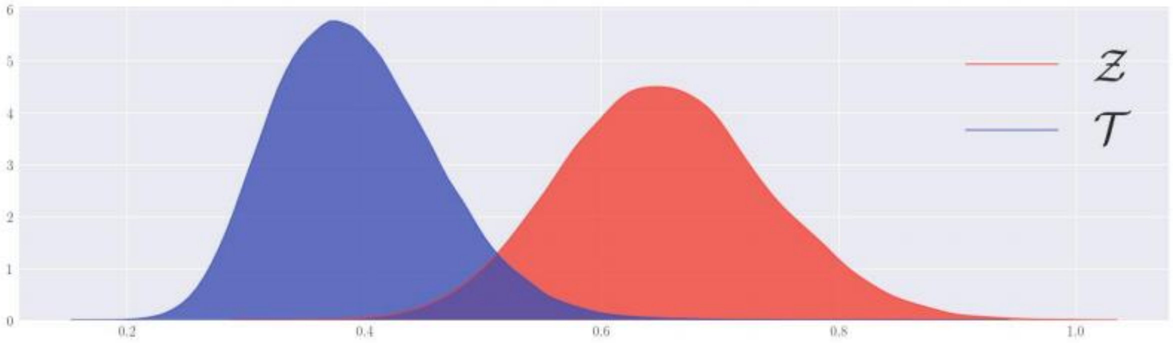
$$\mathcal{T} = \{d(\phi_x^{\mathcal{D}_i}, \phi_x^{\mathcal{D}_j}) \mid f_{\mathcal{D}_i}(x) = y \wedge f_{\mathcal{D}_j}(x) = y\}$$

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

RELATIVE CONSISTENCY

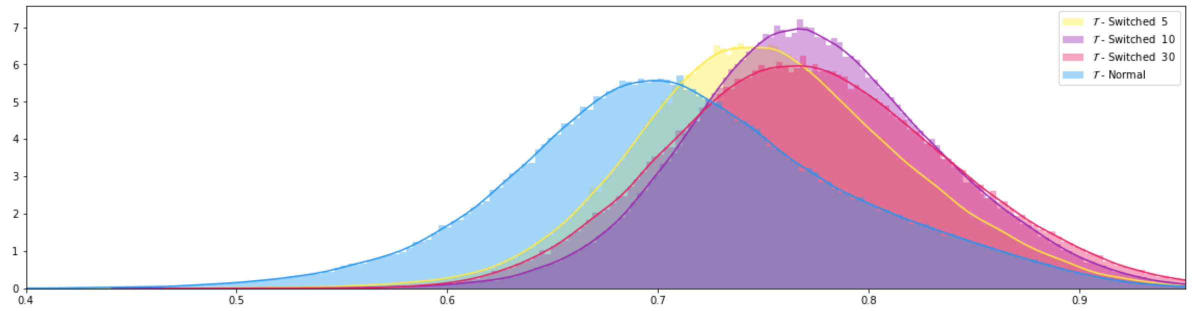
$$ReCo = \max_{\gamma}\ TPR(\gamma) + TNR(\gamma) - 1$$





MEAN GENERALIZABILITY

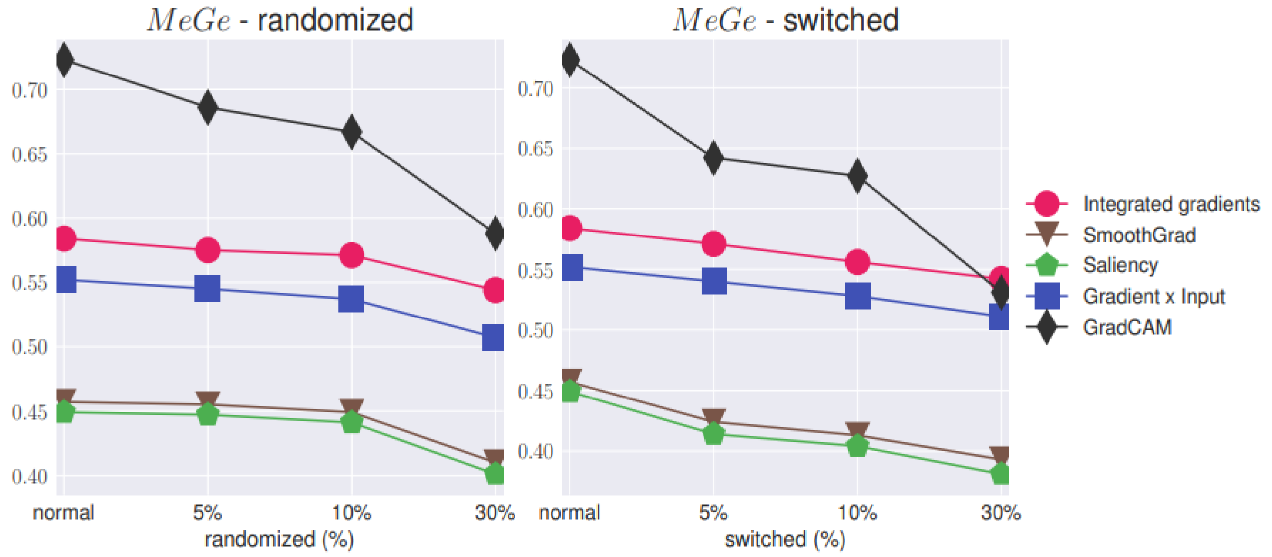
$$MeGe = \frac{1}{1+\mathbb{E}[\mathcal{T}]}$$



EVALUATE METRIC IN FRONT OF MODELS DEGRADATIONS

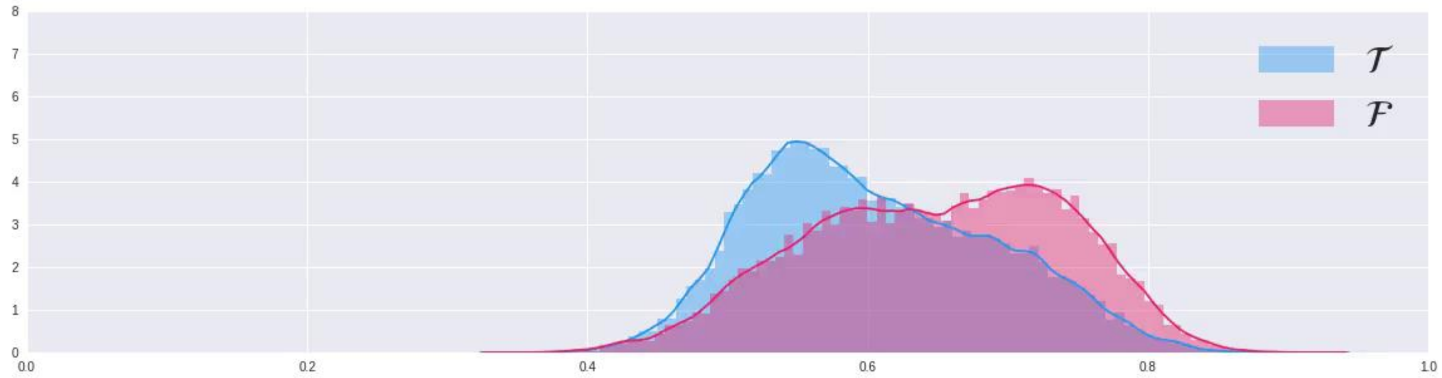
ReCo scores for normally trained ResNet-18

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



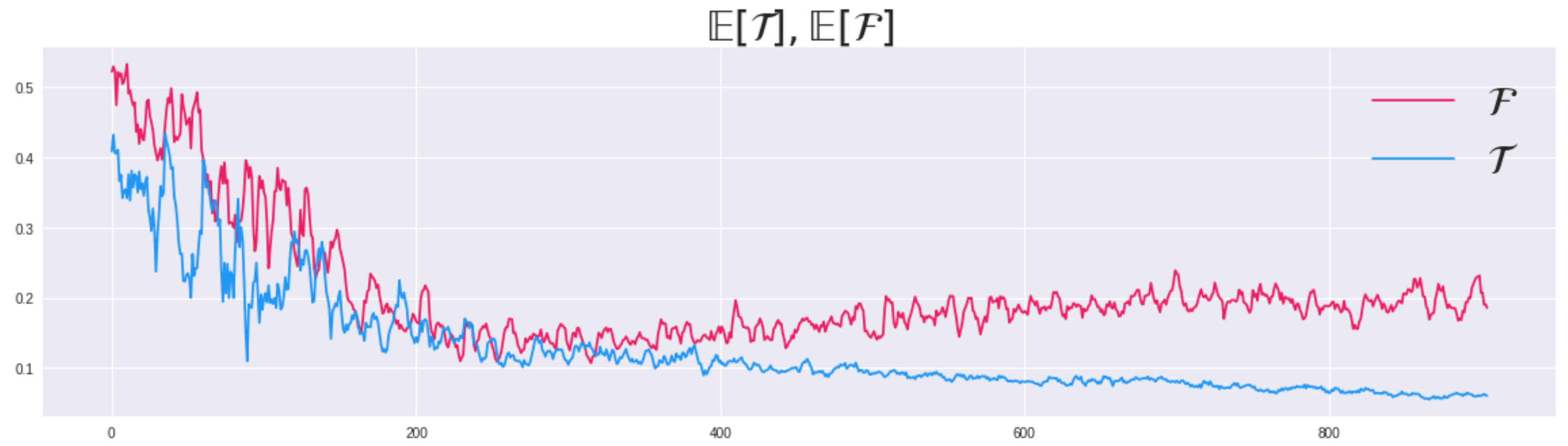
OBSERVATIONS

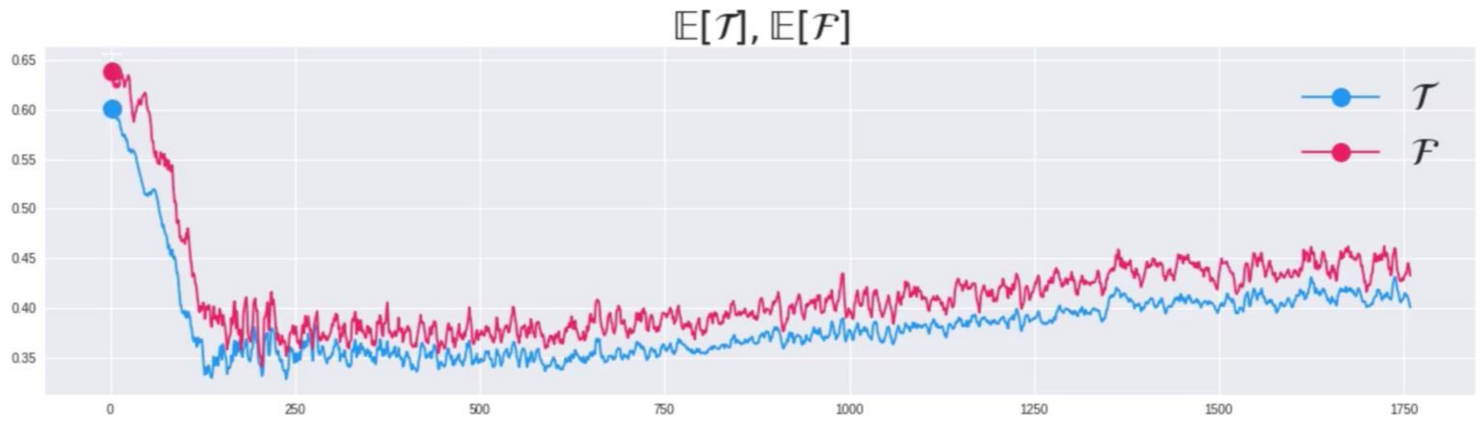
Fashion MNIST



OBSERVATIONS

Fashion MNIST

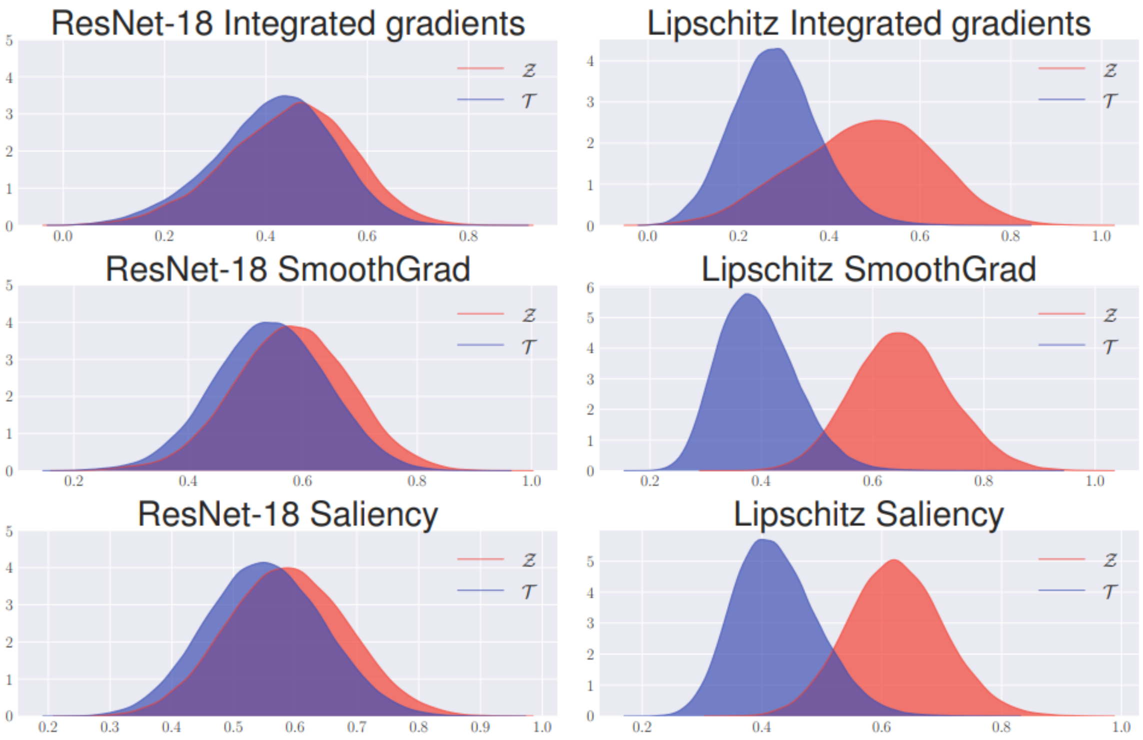


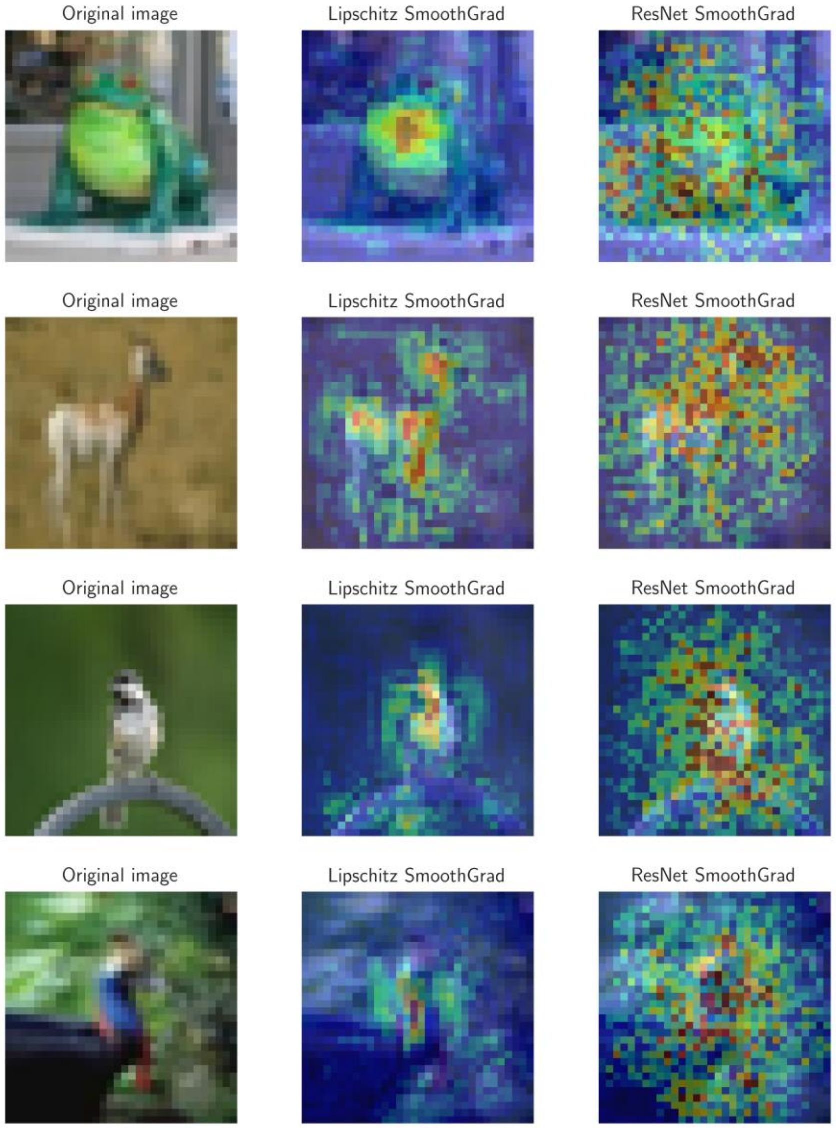


LIPSCHITZ

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

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





Preprint:

https://arxiv.org/abs/2009.04521

Generalization & Consistency Metrics from Deep Neural Network Explanations

Metric team [1 2]

Abstract

The adoption of machine learning in critical contexts requires a reliable explanation of why the algorithm makes certain predictions. To address this issue, many methods have been proposed to explain the choices of these black box models. Despite the choice of these many methods, little effort has been made to ensure that the explanations produced are objectively relevant. While it is possible to establish a number of desirable properties, it is more difficult to associate an evaluation with these properties, i.e. an objective metric to quantify them. As a result, no metrics are actually associated with the properties of consistency and generalization of explanations. We are introducing a new procedure to compute two new metrics, Relative Consistency $ReCo$ and Mean Generalization $MeGe$, respectively for consistency and generalization. Our results on several image classification datasets, and on several progressively degraded models allow us to validate empirically the reliability of these metrics, and reflect qualitative observations reported in previous works. We then put these metrics into practice by applying them to different families of models, revealing an interesting link between gradient-based explanations methods and Lipschitz networks.

1. Introduction

Machine learning techniques such as deep neural networks have become indispensable in multiple domains such as image classification, language processing and speech recognition. These techniques have achieved excellent predictive capability, allowing them to match human performance in many cases.

*Equal contribution [1]IRT Saint Exupery, DEEL, Toulouse, France [2]IRIT, Université de Toulouse, CNRS. Correspondence to: Thomas Fel <thomas.fel@irt-saintexupery.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

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 (Adebayo et al., 2018; Ghorbani 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 ensure 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 could then 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 that the shift does not come from a distribution shift(Rieger & Hansen, 2020; Hooker 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 (Ancona 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

08/04/2021

DEEL
DEpendable & Explainable Learning

MERCI POUR VOTRE ATTENTION

IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE