IA de Confiance pour le Train Autonome

Le besoin de Certification et le challenge de l'Explicabilité

Laurent Gardès & Thomas FEL

Intervenants

Laurent Gardès
Responsable du Plateau IA chez SNCF Innovation & Recherche

Thomas FEL
Doctorant (ANITI, Brown University) sur les métriques d'explicabilité chez SNCF Innovation & Recherche

Plan

Certifier le Train Autonome

- Le cas d’usage
- ML et Train Autonome
- Le besoin en Certification

Le Challenge de l’Explicabilité

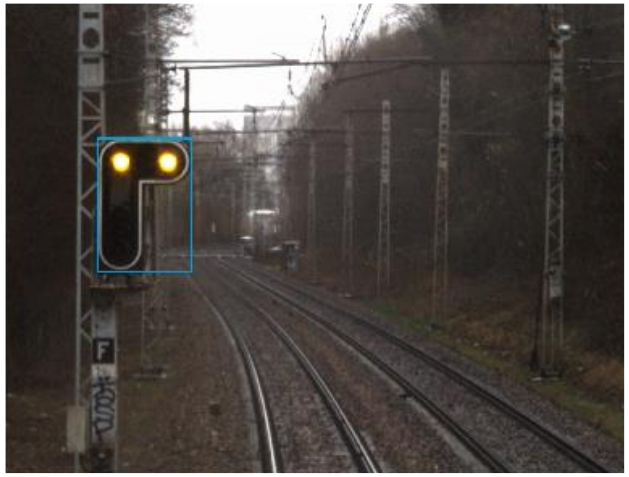
- Aperçu des outils
- Enjeux et problèmes liés aux outils
- Le besoin de Métriques

L’IA de Confiance
Certifier le Train Autonome

SNCF DEEL DEpendable & Explainable Learning IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE BROWN

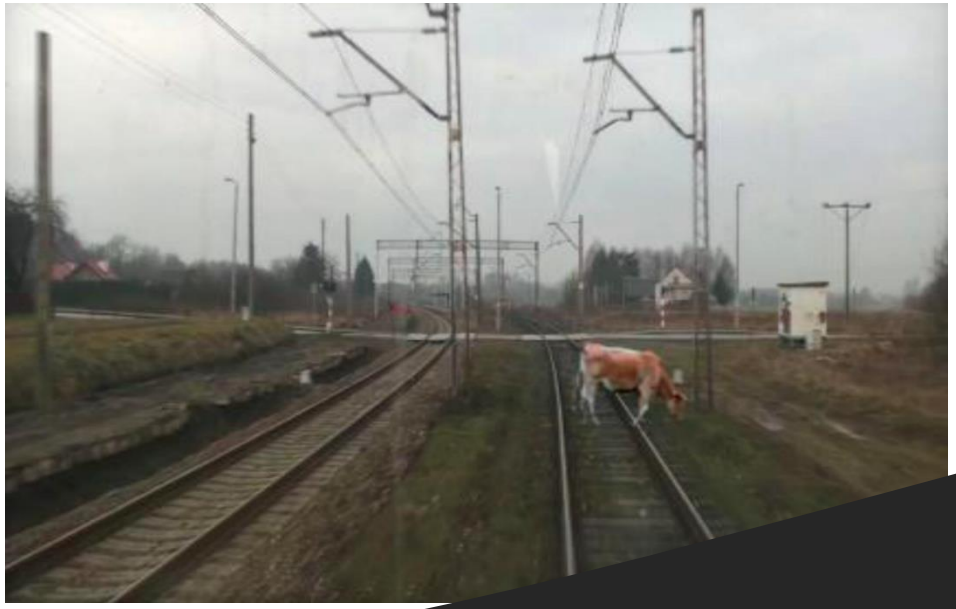
Les Cas d'Usage

TRAIN AUTONOME



LECTURE DE LA SIGNALISATION

+ Conserver le réseau actuel sans modification de l'infrastructure
+ Assurer les fonctions du conducteur avec au moins le même niveau de qualité



DETECTION D'OBSTACLES ET SURVEILLANCE DE L'ENVIRONNEMENT

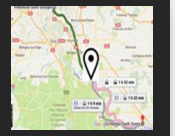
+ Détection d'obstacles dans le gabarit
+ Observer les alentours, les trains croiseurs, etc.

Lecture de la Signalisation

Cartographie du réseau

Chaque signal est cartographié précisément

Localisation du train



Position précise en temps-réel du train

Zone d'intérêt dans l'image

Sélection dans l'image de la zone d'intérêt avec une bounding box

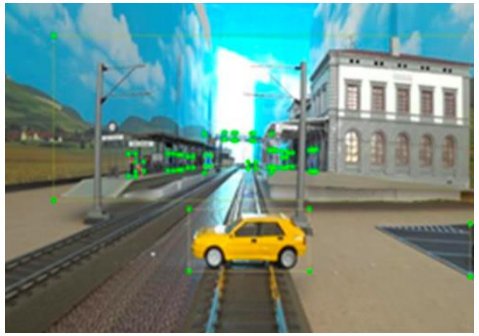
- Moins de pixels à analyser afin d'augmenter les performances de la reconnaissance (évite aussi d'interprèter un mauvais signal)
- Cartographie embarquée: Le système sait ce qu'il doit lire et sait s'il a manqué un signal

Zone d'intérêt

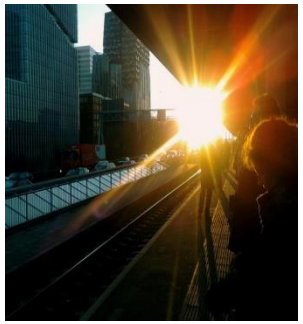
DATASET D'APPRENTISSAGE ET DE TEST

Essais statiques

Essais dynamiques
(annotation manuelle)



Virtualisation



Tests aux limites
(dont usage de GAN)

Data Augmentation

DIFFÉRENTES APPROCHES ALGORITHMIQUES

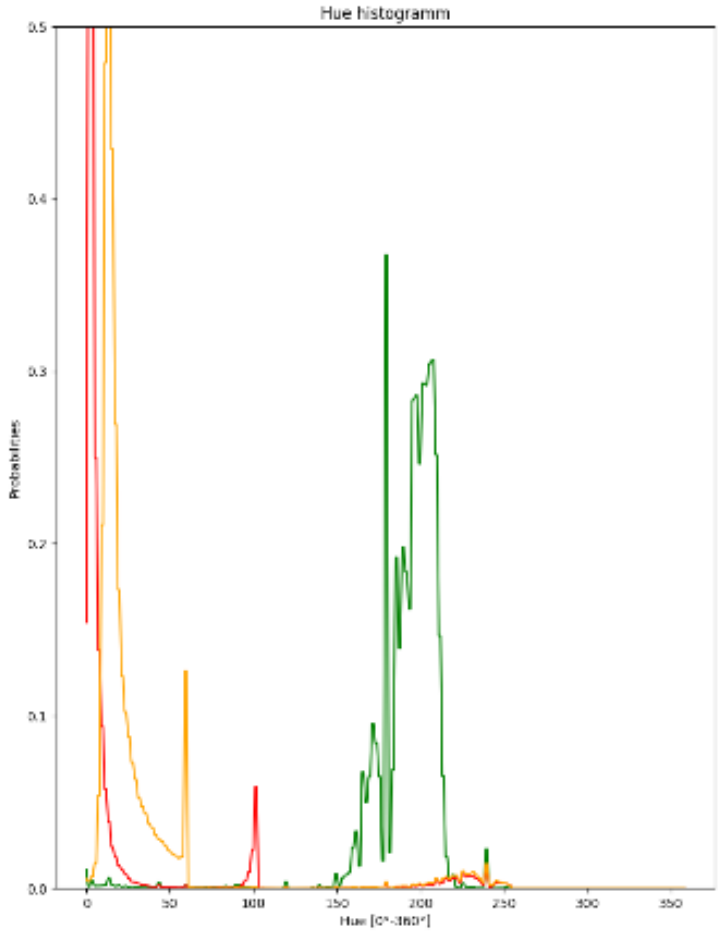
COMPUTER VISION CLASSIQUE

+ Matching d'histogramme
+ Performance qui diminue rapidement avec la distance aux feux

DEEP LEARNING

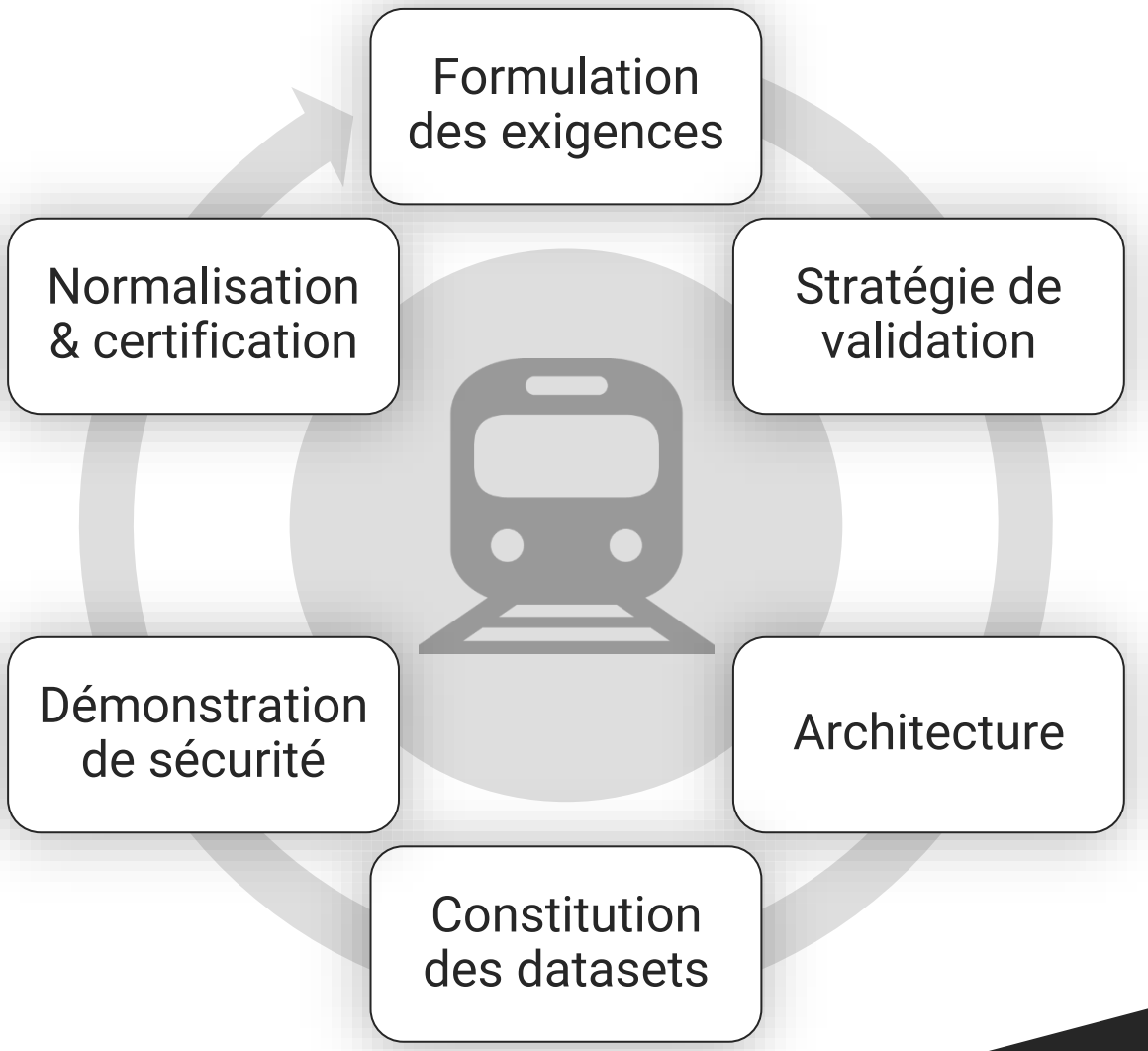
+ CNN et Séquence d'images
+ Performance autour de 99%

MAIS ON EST LOIN DES EXIGENCES EN MATIERE DE SAFETY



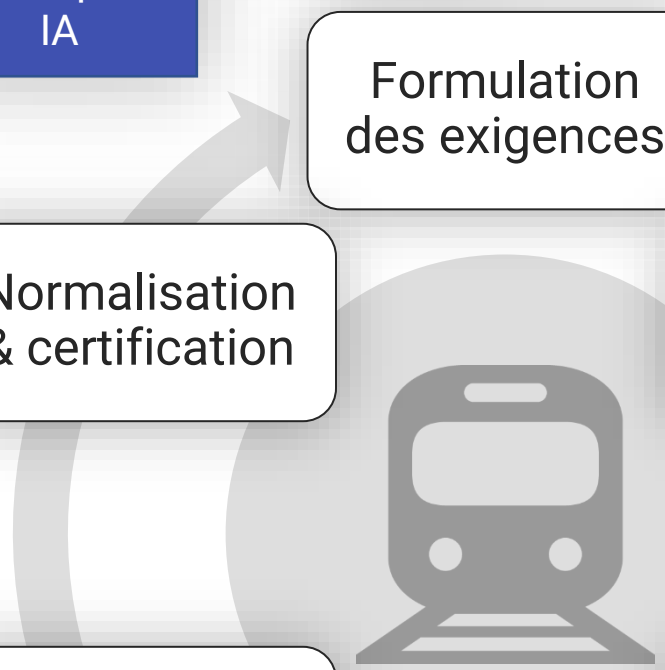
IMPACTS DE L'IA

Un changement de paradigme est nécessaire pour bénéficier des performances de l'IA dans un système critique



IMPACTS DE L'IA

Un changement de paradigme est nécessaire pour bénéficier des performances de l'IA dans un système critique



Lien THR / Métriques IA

Formulation des exigences

Adaptée à l'IA

Nouvelles normes intégrant l'IA

Normalisation & certification

Stratégie de validation

Réduire le niveau d'exigence sur l'IA

Architecture

Démonstration de sécurité

Approche probabiliste ? Méthodes formelles ?

Constitution des datasets

Volumétrie, représentativité, etc.

SNCF DEEL DEpendable & Explainable Learning IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE BROWN

L'IA BOITE NOIRE

Incompréhensible ?

Le grand nombre d'éléments et leurs interconnexions rendent l'interprétation directe d'un modèle de deep learning impossible (analogie : regarder un cerveau humain pour y « voir » une pensée)

Un réseau VGG 16 doit apprendre 138 millions de paramètres

Chaque situation de la vie réelle est **nouvelle** pour l'algorithme

Imprévisible ?

L'espace d'entrée (qui est une réduction du monde réel) est de dimension quasi-infinie

Injuste ?

Le modèle ne fait que reproduire les biais des données d'apprentissage

Nous avons besoin d'outils d'**explicabilité, de robustesse et assurant la représentativité des données (en limitant les biais)**

SNCF DEEL DEpendable & Explainable Learning IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE BROWN

L’IA de Confiance
Le Challenge de l’Explicabilité

Overview – Méthodes d’Attribution – Le besoin de Métriques

SNCF DEEL DEpendable & Explainable Learning IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE BROWN

Overview

Pourquoi avons-nous besoin d'explications ?

Renforcer la confiance dans les prédictions du modèle[3][4][5]

Élucider les aspects importants des modèles appris[4]

Contribuer à satisfaire aux exigences réglementaires et au processus de certification[1]

Révéler les biais ou autres effets non intentionnels appris par un modèle[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
[5] DEEL White Paper

Overview

Un Défi Conceptuel

> “An Explanation is a set of statements [...] which clarifies the cause, the context and consequences of those fact. [...] The component of an explanation can be implicit and interwoven with one another“
>
> *Jess DRAKE, LOGIC*

- Une explication fournit une information
- Une explication dépend de la connaissance du domaine
- Une explication aide à compléter la connaissance du domaine

Overview

Un Défi Technique

Model

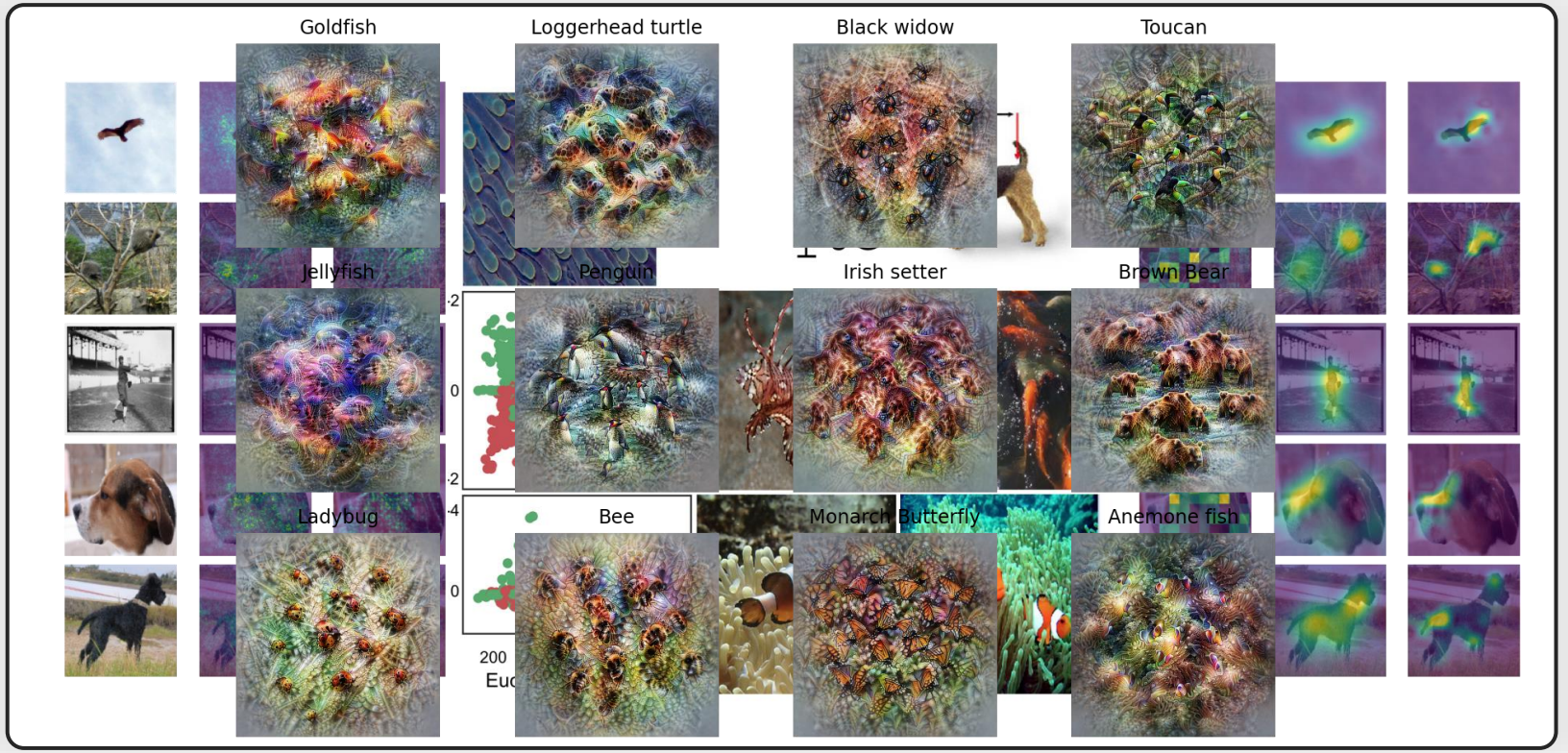
Feature Viz,
Concept Activation Vector
...

Predictions

Feature Attribution
Feature Inversion
...

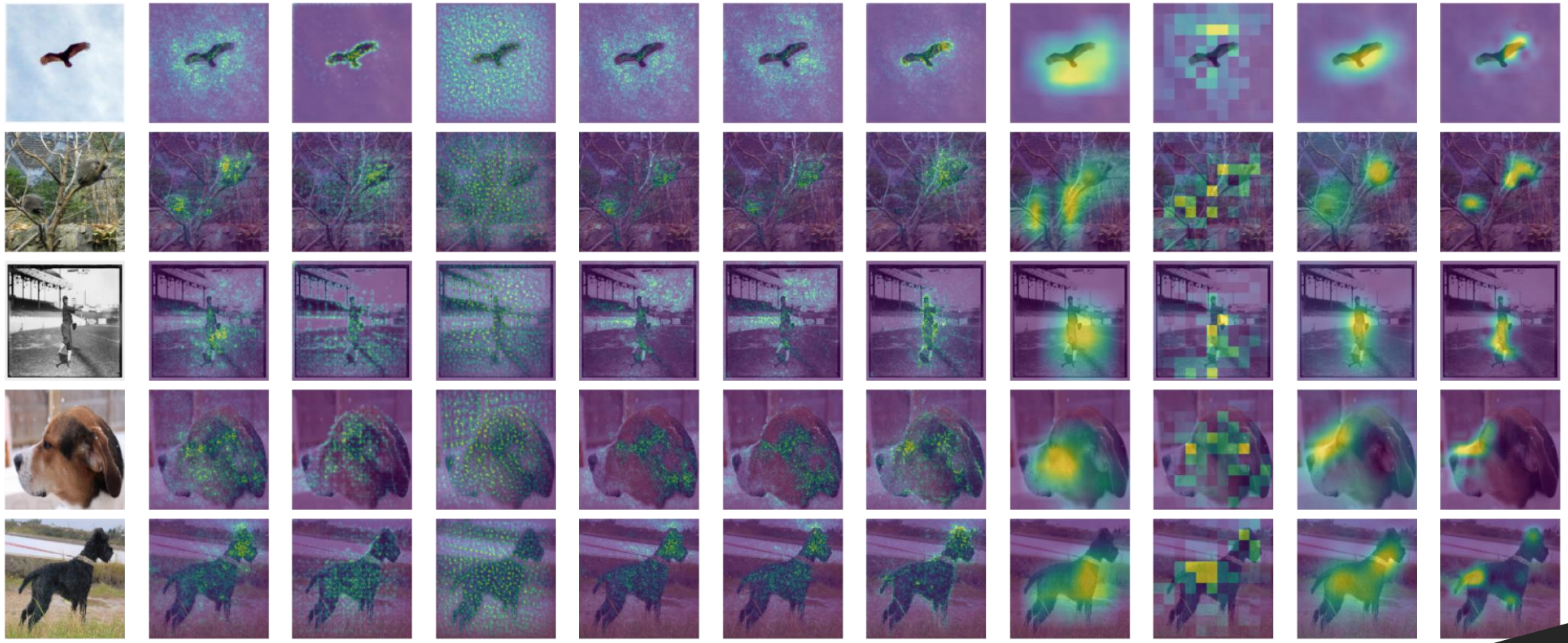
Data

Nearest Neighbourhood
Influence Function
...



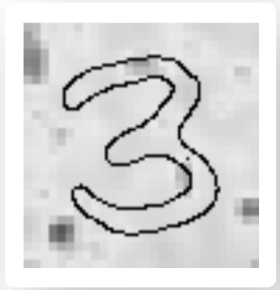
Les Méthodes d'Attribution

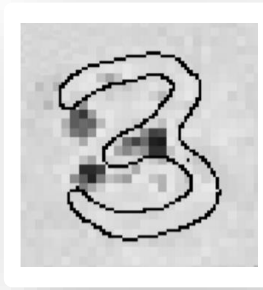
Comment savoir quelle méthode choisir ?

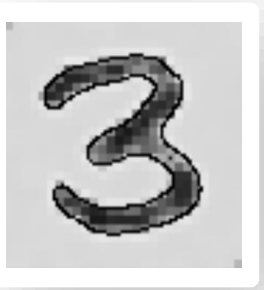


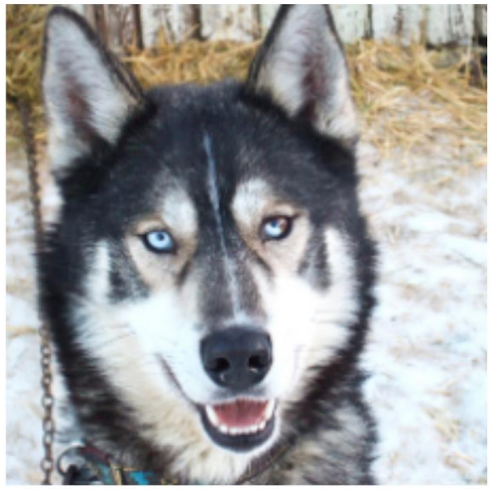
Les besoin de Métriques

Qu'est-ce qu'une bonne explication ?



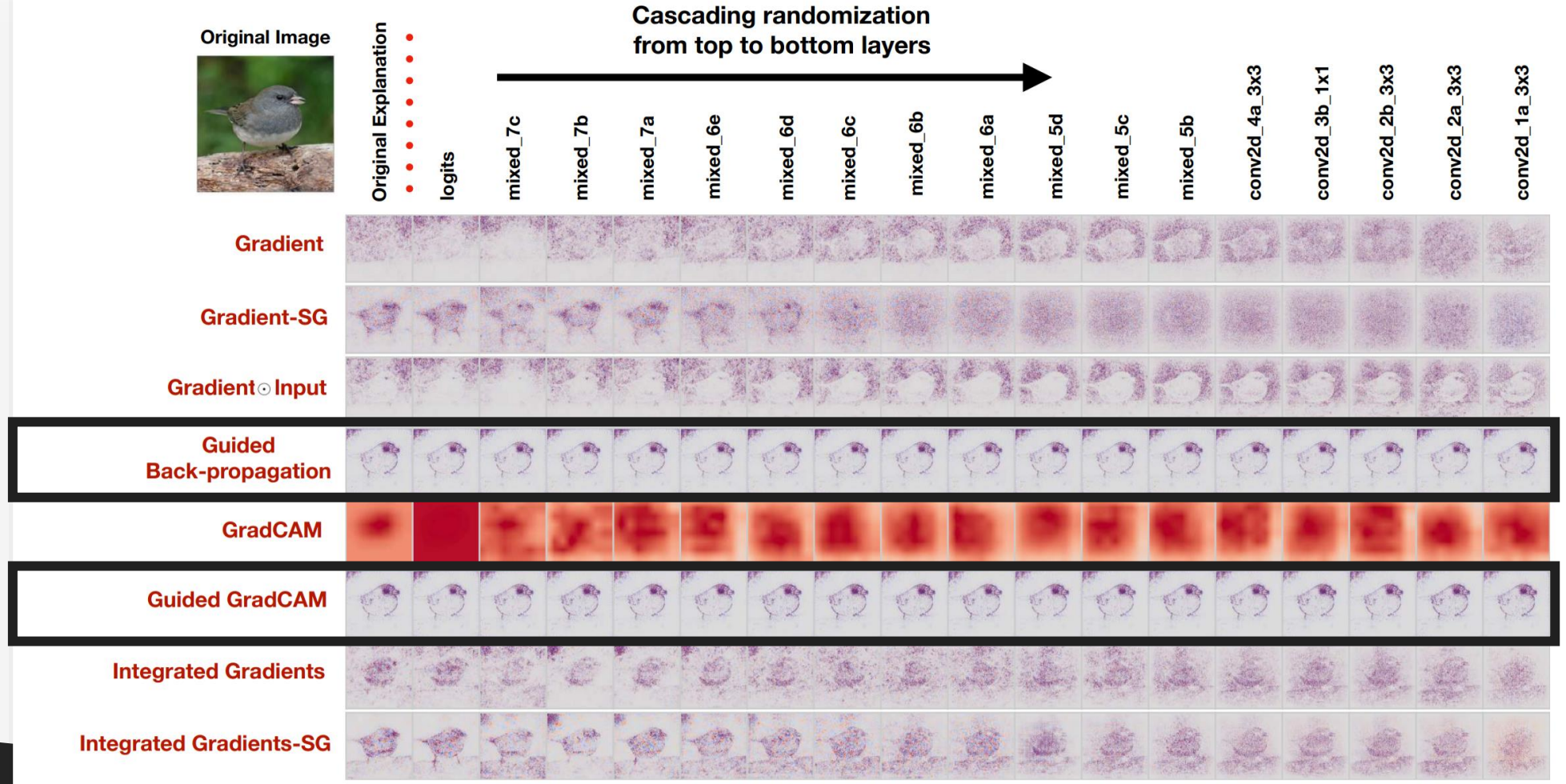






(a) Husky classified as wolf

(b) Explanation


Original Image
Cascading randomization
from top to bottom layers
Original Explanation
logits
mixed_7c
mixed_7b
mixed_7a
mixed_6e
mixed_6d
mixed_6c
mixed_6b
mixed_6a
mixed_5d
mixed_5c
mixed_5b
conv2d_4a_3x3
conv2d_3b_1x1
conv2d_2b_3x3
conv2d_2a_3x3
conv2d_1a_3x3
Gradient
Gradient-SG
Gradient ⊙ Input
Guided
Back-propagation
GradCAM
Guided GradCAM
Integrated Gradients
Integrated Gradients-SG

Confirmation bias.

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

Les besoin de Métriques

Les méthodes d'attribution peuvent être manipulées

Fairwashing Explanations with Off-Manifold Detergent

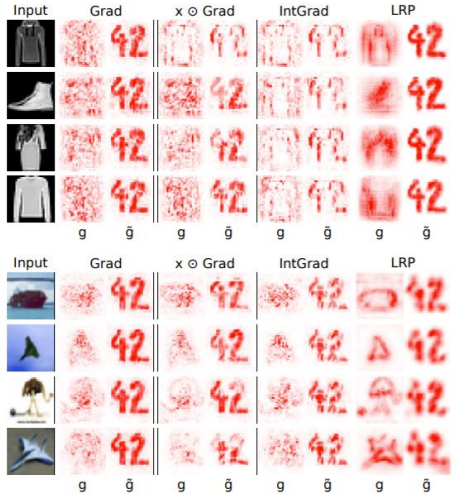
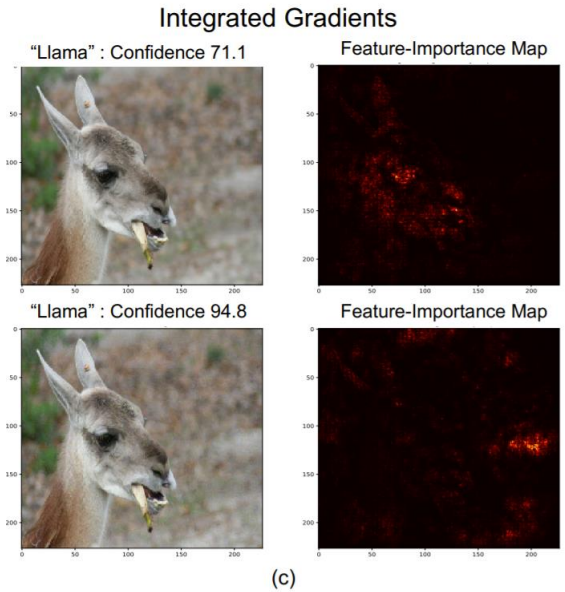


Figure 2. Example explanations from the original model g (left) and the manipulated model $\bar{g}$ (right). Images from the test sets of FashionMNIST (**top**) and CIFAR10 (**bottom**).

Interpretation of Neural Networks is Fragile



(c)

Interpretable Deep Learning under Fire

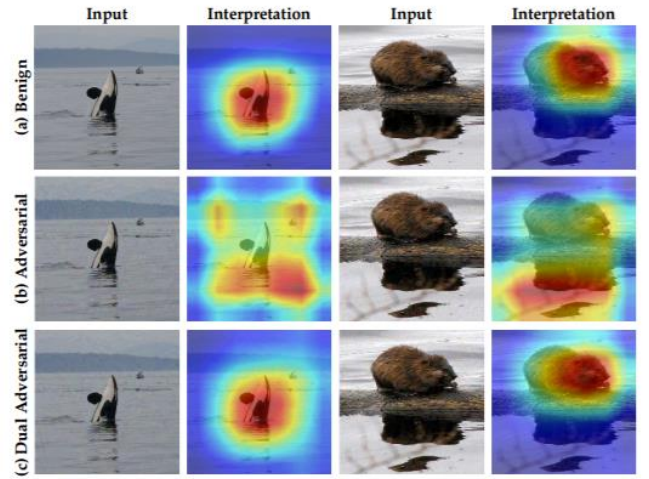


Figure 1: Sample (a) benign, (b) regular adversarial, and (c) dual adversarial inputs and interpretations on ResNet [22] (classifier) and CAM [64] (interpreter).

Les besoin de Métriques

Les propriétés d'une bonne explication

Fidelity	Mon explication reflète-t-elle le comportement de mon modèle ?
Representativity	Combien de phénomènes mon explication couvre-t-elle ?
Comprehensibility	Mon explication est-elle simple et sans ambiguïté ?
Consistency	La mesure dans laquelle des explications similaires sont générées à partir de différents modèles formés sur la même tâche.
Stability	Mon explication reste-t-elle la même sous une transformation sémantiquement invariante ?
Novelty	Mon explication reflète-t-elle le fait que l'instance expliquée provient d'une nouvelle région, non contenue ou non représentée dans l'ensemble d'apprentissage ?

Les besoin de Métriques

Les métriques de Fidélité

Baseline biased

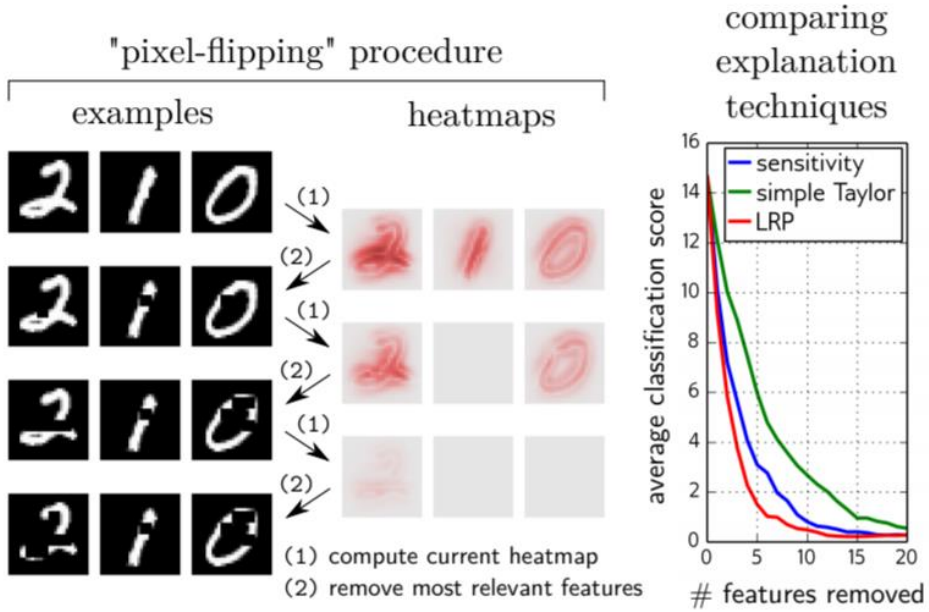


Figure 8: Illustration of the "pixel-flipping" procedure. At each step, the heatmap is used to determine which region to remove (by setting it to black), and the classification score is recorded.

$$F = \underset{S}{\text{corr}} \left(\sum_{i \in S} g(f, x)_i, f(x) - f(x_{[x_i = \bar{x}_i, i \in S]}) \right)$$

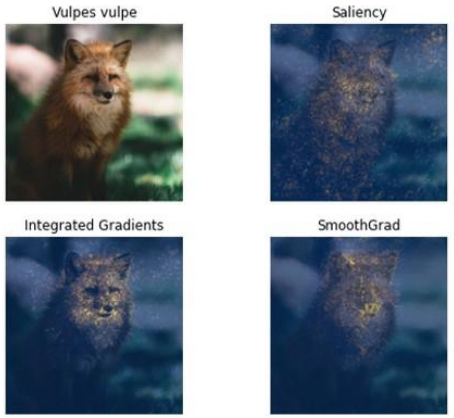
Les besoin de Métriques

Prochaines étapes

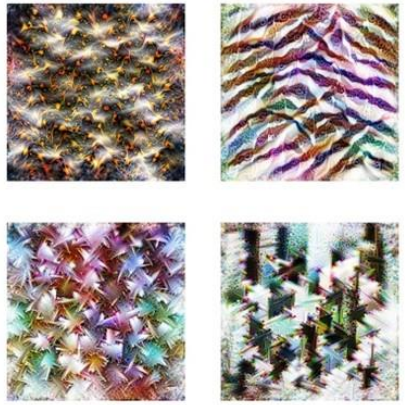
- Nous devons nous assurer que les humains comprennent ces méthodes
- Les méthodes doivent faire l'objet d'étude plus théorique
- Nous avons besoin de plus de Métriques
- L'explication peut / doit être pensé en amont (explanation by design)

XPLIQUE

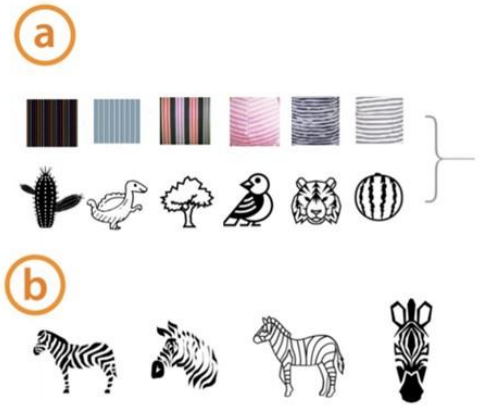
Vision Explainability Toolbox for Tensorflow



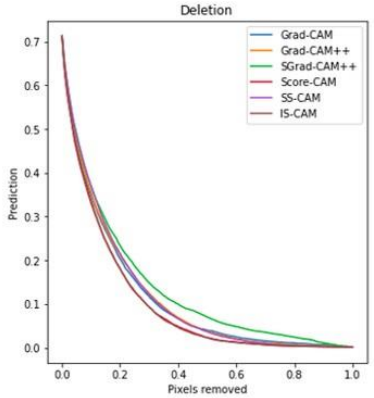
Feature Attribution



Feature Visualization



Concept Extraction



Metrics

https://github.com/deel-ai/xplique

SNCF DEEL DEpendable & Explainable Learning IVADO IRT SAINT EXUPERY CRIAQ ANITI ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE BROWN

Merci de votre attention, quelles questions avez-vous ?

Laurent Gardès, Thomas FEL
SNCF Innovation & Recherche