Explicabilité et transparence en apprentissage basé sur les statistiques

Philippe Preux
philippe.preux@univ-lille.fr

Extracting Provably Correct Rules from Artificial Neural Networks

Sebastian B. Thrun
University of Bonn
Dept. of Computer Science III
Römerstr. 164, D-5300 Bonn 1, Germany
E-mail: thrun@cs.uni-bonn.de thrun@cmu.edu
Phone: +49-228-550-260 FAX: +49-228-550-382

Abstract

Although connectionist learning procedures have been applied successfully to a variety of real-world scenarios, artificial neural networks have often been criticized for exhibiting a low degree of comprehensibility. Mechanisms that automatically compile neural networks into symbolic rules offer a promising perspective to overcome this practical shortcoming of neural network representations.

This paper describes an approach to neural network rule extraction based on *Va-*

Extracting Provably Correct Rules from Artificial Neural Networks

Sebastian B. Thrun

University of Bonn
Dept. of Computer Science III
Römerstr. 164, D-5300 Bonn 1, Germany
E-mail: thrun@cs.uni-bonn.de thrun@cmu.edu
Phone: +49-228-550-260 FAX: +49-228-550-382

Publié à NIPS 1995

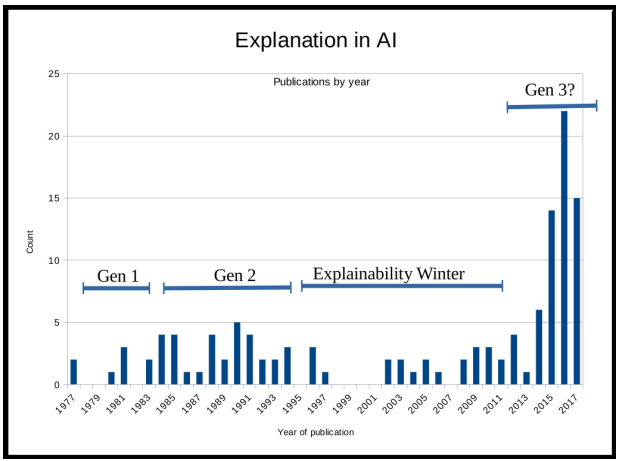
Abstract

suite d'un papier à ICML 1993, NIPS 1992, ...

Although connectionist learning procedures have been applied successfully to a variety of real-world scenarios, artificial neural networks have often been criticized for exhibiting a low degree of comprehensibility. Mechanisms that automatically compile neural networks into symbolic rules offer a promising perspective to overcome this practical shortcoming of neural network representations.

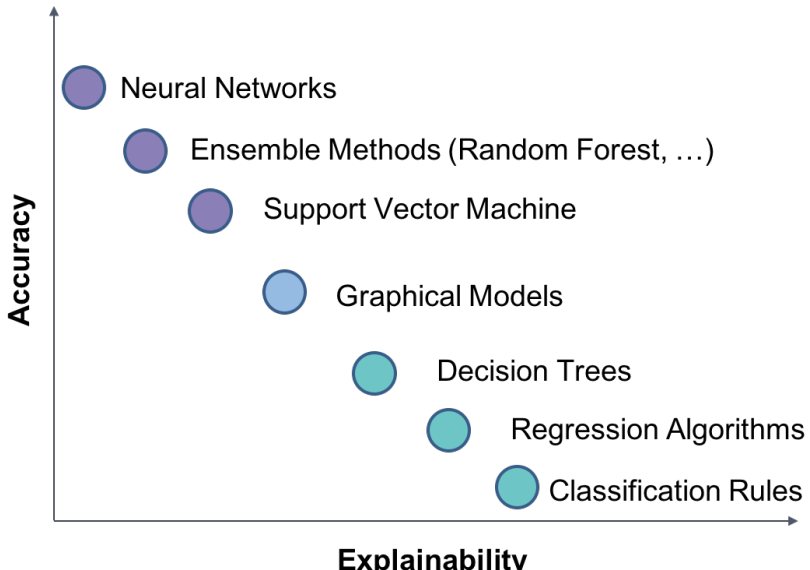
This paper describes an approach to neural network rule extraction based on *Va-*

Explicabilité : une histoire ancienne



"Histogram of the number of publications per year identified in our literature review that are relevant to explanation in intelligent systems (AI, machine learning, and related fields)." Source : (1902.01876)

Compromis précision *vs.* transparence



(⚠ Cette figure n'est pas très précise, mais donne l'idée à retenir.)

Pourquoi s'intéresser à la question ?

- RGPD

Pourquoi s'intéresser à la question ?

- RGPD
- sensibilité de certains domaines d'application : santé, véhicule autonome, prêts bancaires, assurances, justice, ...

Pourquoi s'intéresser à la question ?

- RGPD
- sensibilité de certains domaines d'application : santé, véhicule autonome, prêts bancaires, assurances, justice, ...
- parce que la question pose des problèmes de recherche intéressants

2 problématiques

Interaction human – machine apprenante :

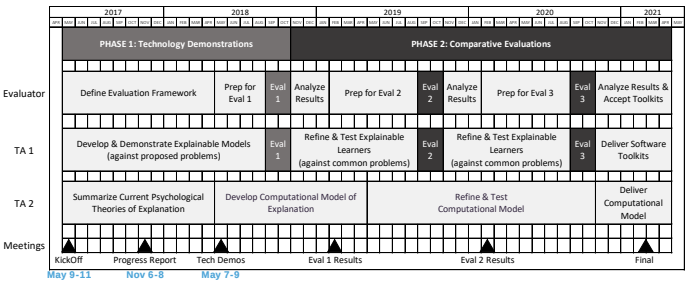
- comprendre les raisons ayant entraîné une certaine décision de la part d'une machine apprenante
- pouvoir exprimer des informations à destination de la machine pour influer sur son apprentissage

Plan

- intro
- projet XAI du DARPA
- projet Hy_AIAI de l'Inria
- exemple de travaux sur l'explication en classification supervisée
- conclusion

XAI, “Xplainable AI”

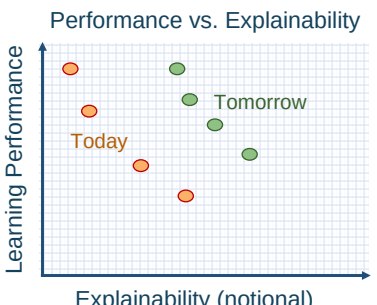
- projet DARPA 2017 – 2021 : appel à soumission de projet en août 2016.
- https://www.darpa.mil/program/explainable-artificial-intelligence



- environ 80 publications en 2017 sur https://www.darpa.mil/opencatalog
- 1 parmi 20 programmes IA du DARPA, budget 2.10^9 US$ sur 5 ans.

Goal: Performance and Explainability

- XAI will create a suite of machine learning techniques that
  - Produce more explainable models, while maintaining a high level of learning performance (e.g., prediction accuracy)
  - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners



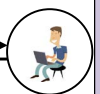
XAI Concept and Technical Approaches

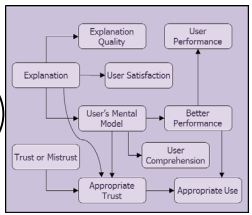
Training Data

New Learning Process

Explainable Model

Explanation Interface





IHMC
Psychological Model of Explanation

UC Berkeley	Deep Learning	Reflexive and Rational
Charles River Analytics	Causal Modeling	Narrative Generation
UCLA	Pattern Theory+	3-Level Explanation
Oregon State	Adaptive Programs	Acceptance Testing
PARC	Cognitive Modeling	Interactive Training
CMU	Explainable RL (XRL)	XRL Interaction
SRI International	Deep Learning	Show and Tell Explanations
Raytheon BBN	Deep Learning	Argumentation and Pedagogy
UT Dallas	Probabilistic Logic	Decision Diagrams
Texas A&M	Mimic Learning	Interactive Visualization
Rutgers	Model Induction	Bayesian Teaching

Explication ?

Le programme XAI mentionne 3 types d'explication, tout en laissant la porte ouverte à d'autres :

- *deep explaination* : développer de nouvelles techniques de DL pour apprendre des features plus explicatifs, des représentations explicables, ou des systèmes générant des explications.
- *interpretable models* : développer de nouvelles techniques de ML qui apprennent des modèles plus structurés, plus interprêtables, ou causaux, *e.g.* listes de règles bayésiennes, *Bayesian Program Learning*, modèles causaux, grammaires stochastiques, ...
- *model induction* : étant donné un modèle en boîte noire, en construire un modèle approximatif et *explainable* (*cf.* RIbeiro *et al.*. KDD 2016])

Hy_AIAI

- "Inria project Lab", 2019 – 2022 (kick-off en juillet 2019)
- réunion de 6 EP : Lacodam, Magnet, Multispeech, Orpailleur, SequeL, Tau.
- apprentissage symbolique et numérique.
- budget $\approx$ 600 k€

Hy_AIAI : cibles

- le système doit comprendre les attentes de l'humain.
  Utiliser un graphe de connaissances lors de l'entraînement d'un RN
- les humains doivent comprendre les réponses du système
  Passer d'une compréhension globale d'un RN à une compréhension lacalisée/focalisée.
- les humains doivent comprendre le fonctionnement interne du système (*debugging*)
  Peut-on trouver ce qui provoque des prédictions erronnées dans un RN entraîné ?
- causalité

Explication classification supervisée d'images

Hendricks *et al.*, *Generating Visual Explanations*, ECCV 2016, 1603.08507

Western Grebe

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Western Grebe* is a waterbird with a yellow pointy beak, white neck and belly, and black back.
Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

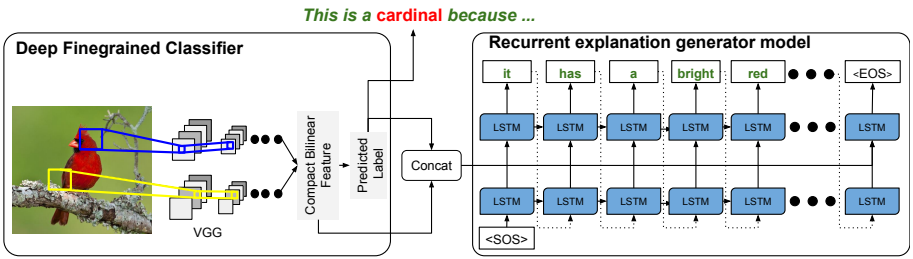
Laysan Albatross

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Objectif : fournir une explication en langage naturel de la raison pour laquelle une donnée est classée d'une certaine façon.

Explication classification supervisée d'images

Hendricks *et al.*, *Generating Visual Explanations*, ECCV 2016, 1603.08507



Encourage la génération d'une phrase qui est discriminante pour la classe prédite de l'objet.

Explication classification supervisée d'images

Hendricks *et al.*, *Generating Visual Explanations*, ECCV 2016, 1603.08507

- basé sur un "Long-Term Recurrent Convolutional Network" (LRCN)
  [Donahue *et al.*, CVPR 2015, 1411.4389]
- qui est une combinaison de :
  - 1 réseau à convolution pour extraire des caractéristiques visuelles
  - 2 réseaux récurrents pour générer une description des caractéristiques visuelles.
    - le premier génère une phrase correspondant à la classe prédite pour l'image considérée
      Entraînement supervisé : 1 exemple = (image, classe, phrase)
    - le second génère une phrase discriminante pour cette classe
      Entraînement par apprentissage par renforcement
- fonction optimisée :

$$\text{Relevance (description)} + \lambda \text{Discrimination (description)}$$

- entrainement sur le jeu de données Caltech UCSD Birds 200-2011 : 200 classes, 11788 images en tout pour chacune desquelles 5 phrases descriptives ont été ajoutées.

This is a pine grosbeak because this bird has a red head and breast with a gray wing and white wing.

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

This is an artic tern because this is a white bird with a black head and orange feet.

Definition:	this bird is **black** with **blue** on its wings and has a long **pointy beak**.
Description:	this bird is **nearly all black** with a short **pointy bill**.
Explanation-Label:	this bird is **nearly all black** with **bright orange eyes**.
Explanation-Dis.:	this is a **black bird** with a **red eye** and a **white beak**.
Explanation:	this is a **black bird** with a **red eye** and a **pointy black beak**.

This is a **Black Billed Cuckoo** *because ...*

Definition:	this bird has a **yellow belly** and a **grey head**.
Description:	this bird has a **yellow belly** and **breast** with a **gray crown** and **green wing**.
Explanation-Label:	this bird has a **yellow belly** and a **grey head** with a **grey throat**.
Explanation-Dis.:	this is a **yellow bird** with a **grey head** and a **small beak**.
Explanation:	this is a **yellow bird** with a **grey head** and a **pointy beak**.

This is a **White Necked Raven** *because ...*

Definition:	this bird is **black in color** with a **black beak** and **black eye rings**.
Description:	this bird is **black** with a **white spot** and has a **long pointy beak**.
Explanation-Label:	this bird is **black** in color with a **black beak** and **black eye rings**.
Explanation-Dis.:	this is a **black** bird with a **white nape** and a **black beak**.
Explanation:	this is a **black** bird with a **white nape** and a **large black beak**.

This is a **Northern Flicker** *because ...*

Definition:	this bird has a **speckled belly and breast** with a **long pointy bill**.
Description:	this bird has a **long pointed bill grey throat** and **spotted black and white mottled crown**.
Explanation-Label:	this bird has a **speckled belly and breast** with a **long pointy bill**.
Explanation-Dis.:	this is a **grey bird** with **black spots** and a **red spotted crown**.
Explanation:	this is a **black and white spotted bird** with a **red nape** and a **long pointed black beak**.

This is a **American Goldfinch** *because ...*

Definition:	this bird has a **yellow crown** a **short and sharp bill** and a **black wing** with a **white breast**.
Description:	this bird has a **black crown** a **yellow bill** and a **yellow belly**.
Explanation-Label:	this bird has a **black crown** a **short orange bill** and a **bright yellow breast and belly**.
Explanation-Dis.:	this is a **yellow bird** with a **black wing** and a **black crown**.
Explanation:	this is a **yellow bird** with a **black and white wing** and an **orange beak**.

This is a **Yellow Breasted Chat** *because ...*

Definition:	this bird has a **yellow belly and breast** with a **white eyebrow** and **gray crown**.
Description:	this bird has a **yellow breast and throat** with a **white belly and abdomen**.
Explanation-Label:	this bird has a **yellow belly and breast** with a **white eyebrow** and **gray crown.**
Explanation-Dis.:	this is a bird with a **yellow belly** and a **grey back and head**.
Explanation:	this is a bird with a **yellow breast** and a **grey head and back**.

This is a **Hooded Merganser** *because ...*

Definition:	this bird has a **black crown** a **white eye** and a **large black bill**.
Description:	this bird has a **brown crown** a **white breast** and a **large wingspan**.
Explanation-Label:	this bird has a **black and white head** with a large **long yellow bill** and **brown tarsus and feet.**
Explanation-Dis.:	this is a **brown bird** with a **white breast** and a **white head**

*This is a **Black-Capped Vireo** because...*

Description: this bird has a white belly and breast black and white wings with a white wingbar.

Explanation-Dis: this is a bird with a white belly yellow wing and a **black head**.

*This is a **Crested Auklet** because...*

Description: this bird is black and white in color with a orange beak and black eye rings.

Explanation-Dis.: this is a black bird with a **white eye** and an orange beak.

*This is a **Green Jay** because...*

Description: this bird has a bright blue crown and a bright yellow throat and breast.

Explanation-Dis.: this is a yellow bird with a **blue head** and a **black throat**.

*This is a **White Pelican** because...*

Description: this bird is white and black in color with a long curved beak and white eye rings.

Explanation: this is a large white bird with a **long neck** and a **large orange beak**.

*This is a **Geococcyx** because...*

Description: this bird has a long black bill a white throat and a brown crown.

Explanation-Dis.: this is a black and white spotted bird with a **long tail feather** and a pointed beak.

*This is a **Cape Glossy Starling** because...*

Description: this bird is blue and black in color with a stubby beak and black eye rings.

Explanation-Dis.: this is a blue bird with a **red eye** and a blue crown.

This is a ***Baltimore Oriole*** *because* this is a small bird with a black head and orange body with black wings and tail.
This is a ***Cliff Swallow*** *because* this bird has a black crown a black throat and a white belly.
This is a ***Painted Bunting*** *because* this is a colorful bird with a red belly green head and a yellow throat.

This is a ***Baltimore Oriole*** *because* this is a small bird with a black head and a small beak.
This is a ***Cliff Swallow*** *because* this bird has a black crown a brown wing and a white breast.
This is a ***Painted Bunting*** *because* this is a small bird with a red belly and a blue head.

This is a ***Baltimore Oriole*** *because* this is a small orange bird with a black head and a small orange beak.
This is a ***Cliff Swallow*** *because* this is a black bird with a red throat and a white belly.
This is a ***Painted Bunting*** *because* this is a colorful bird with a red belly green head and a yellow throat.

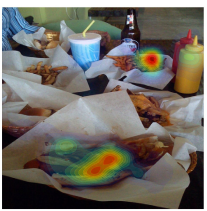
Explication multimodale

Park *et al..*, *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence*, CVPR 2018, 1802.08129

Q: Is this a healthy meal? Textual Justification Visual Pointing

A: No

...because it is a hot dog with a lot of toppings.



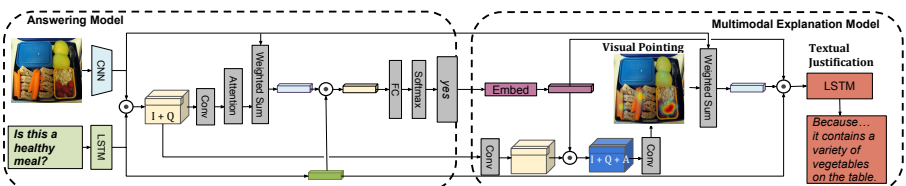
A: Yes

...because it contains a variety of vegetables on the table.

Explication multimodale

Pointing and Justification Model (PJ-X)

- objectif : apprendre à pointer une explication sur une image et à expliquer en langue naturelle.
- tâche : VQA, reconnaissance d'activité.
-

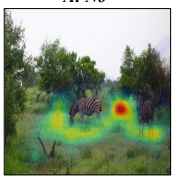


Explication multimodale

Park *et al..*, *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence*, CVPR 2018, 1802.08129

Q: Is this a zoo?

A: No



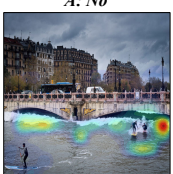
... because the zebras are standing in a green field.

A: Yes

... because there are animals in an enclosure.

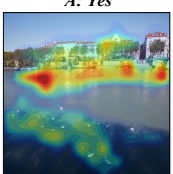
Q: Is the water calm?

A: No



... because there are waves and foam.

A: Yes



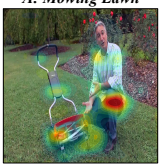
... because there are no waves and you can see the reflection of the sun.

Explication multimodale

Park *et al..*, *Multimodal Explanations: Justifying Decisions and Pointing to the Evidence*, CVPR 2018, 1802.08129

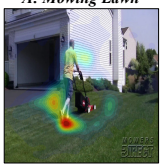
The activity is

A: Mowing Lawn



... because he is kneeling in the grass next to a lawn mower.

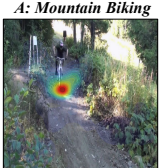
A: Mowing Lawn



... because he is pushing a lawn mower over a grassy lawn.

The activity is

A: Mountain Biking



... because he is riding a bicycle down a mountain path in a mountainous area.

A: Road Biking

... because he is wearing a cycling uniform and riding a bicycle down the road.

Un peu de lecture sur le sujet

- sur XAI : `https://www.darpa.mil/program/explainable-artificial-intelligence`
- *Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI*, Arxiv:1902.01876
- *Transparency in Algorithmic Decision Making, ERCIM News*, **116**, Jan. 2019.
- High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, avril 2019.