Interprétabilité des réseaux de neurones

Master SETI

Julien Girard-Satabin (CEA LIST) : julien.girard2@cea.fr

16 février 2023

Préliminaires

Vous et moi

Moi :

1. ingénieur-chercheur au CEA LIST, docteur en informatique
2. s'intéresse à l'intelligence artificielle de confiance (vérification formelle, interprétabilité)
3. citoyen informés

Vous :

1. élèves niveau M2 du master SETI
2. futurs utilisateurs informés ou développeurs d'IA
3. citoyens informés

Nos objectifs pour ce cours

Moi :

1. susciter l'intérêt sur un domaine de recherche que j'estime passionnant
2. informer sur des techniques importantes
3. recruter !

Vous :

1. obtenir des connaissances sur la notion d'explication, et son instanciation technique sur l'IA
2. connaître les limitations des techniques actuelles d'interprétabilité

Sujets abordés dans ce cours

1. interprétabilité et explication des programmes à base d'IA
2. apprentissage supervisé
3. focus sur la vision par ordinateur, un peu d'analyse de texte
4. quasi exclusivement les réseaux de neurones

Sujets non abordés dans ce cours

1. apprentissage non-supervisé
2. apprentissage par renforcement
3. état de l'art de 2023

Matériel et remerciements

1. Mon collègue Romain Xu-Darme
2. Ce survey [Nau+22]
3. Ce livre [Mol22]

Expliquer : pourquoi, comment ?

Définitions

Explication

« Une explication est une présentation de tout ou partie du fonctionnement d'un programme dans des termes compréhensibles par un humain » [Nau+22]

Définitions

Interprétation

L'interprétation est le processus d'ajustement d'un fait perçu par une personne avec sa représentation mentale de la situation

L'interprétation est une étape du processus d'explication

Définitions

Interprétation

L'interprétation est le processus d'ajustement d'un fait perçu par une personne avec sa représentation mentale de la situation

L'interprétation est une étape du processus d'explication

Plusieurs personnes <=> plusieurs modèles mentaux

Des nuances dans l'interprétation

Que constitue une « bonne » explication, en psychologie et philosophie ([Mil19]) ?

1. explication *contrastive* : pourquoi P plutôt que Q ?
2. explication comme *un processus social* : A explique P à B
3. les explications plus *générales* (qui expliquent plus de choses), plus *simples* (qui citent moins de cause), et *cohérentes* (qui se rapportent à une connaissance précédente) sont plus facilement intégrées

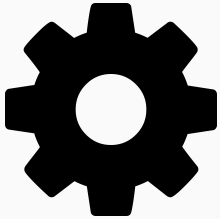
Des nuances dans l'interprétation

Que constitue une « bonne » explication, en psychologie et philosophie ([Mil19]) ?

1. explication *contrastive* : pourquoi P plutôt que Q ?
2. explication comme *un processus social* : A explique P à B
3. les explications plus *générales* (qui expliquent plus de choses), plus *simples* (qui citent moins de cause), et *cohérentes* (qui se rapportent à une connaissance précédente) sont plus facilement intégrées

Il est *nécessaire* que l'explication soit basée sur des faits démontrables, mais ce n'est pas le critère préférentiel

Pourquoi expliquer ?



impots.gouv.fr parcoursup Entrez dans l'enseignement supérieur

Savoir ce que le programme prédit n'est souvent pas suffisant dans un cadre réel

Pourquoi ça importe

1. debugging et audit
2. transparence envers l'utilisateur, acceptation sociale
3. respect du droit

Un exemple légal

RGPD article 13 alinea f

[...] le responsable du traitement fournit à la personne concernée, au moment où les données à caractère personnel sont obtenues, les informations complémentaires suivantes qui sont nécessaires pour garantir un traitement équitable et transparent : [...] l'existence d'une prise de décision automatisée, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, et, au moins en pareils cas, **des informations utiles concernant la logique sous-jacente**, [...]

Considéré comme un « droit à l'explication » selon [SP17]

Un mauvais exemple : COMPAS

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

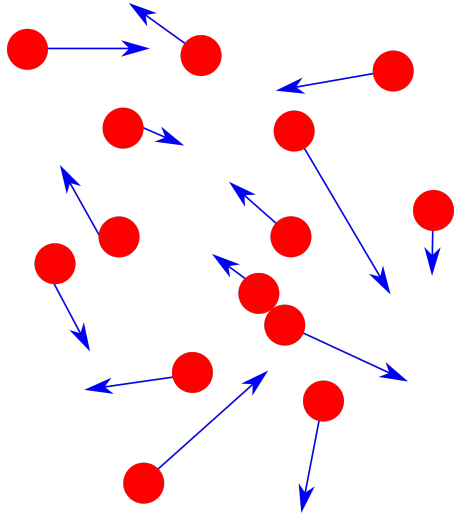
Un mauvais exemple : COMPAS

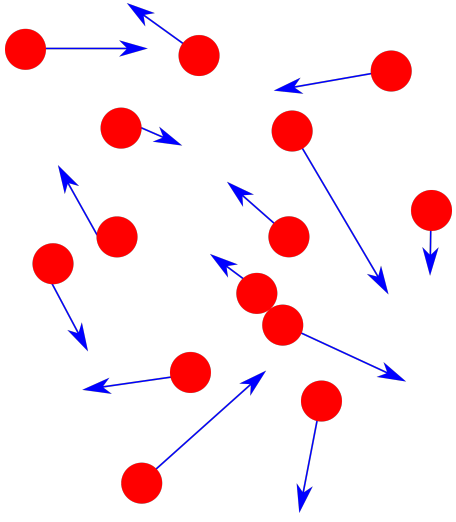
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

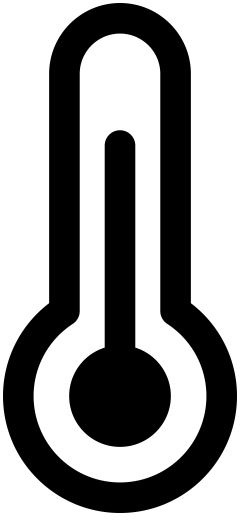
sur quelles prémisses repose un résultat ? Ici, un programme qui perpétue des biais racistes en reproduisant et amplifiant le passé

À propos du terme « black-box »

On sait ce que c'est qu'un réseau de neurones, les principes fondamentaux sont compris







Les différentes approches techniques de l’explicabilité

Une proposition de taxonomie

1. post-hoc
2. par construction

La plupart des techniques présentées proposent d'expliquer *une prédiction particulière du modèle*

Une explication *globale* est compliquée à obtenir

Post-hoc

post-hoc

Une méthode *post-hoc* cherche à interpréter un programme une fois sa conception terminée

Exemple de techniques : attribution de caractéristiques, cartes de chaleurs

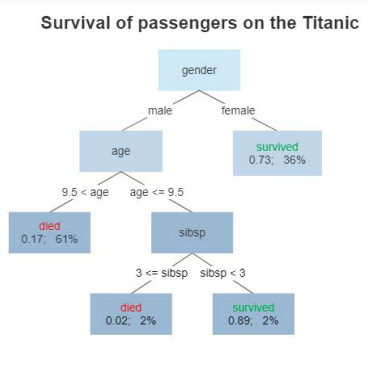
Par construction

Par construction

Concevoir un programme de sorte à faciliter son interprétabilité

Exemples de techniques : arbres de décision, régressions, systèmes experts, modèles d'attention, approches par prototypes

Arbres de décision



De Wikipedia `https://en.wikipedia.org/wiki/Decision_tree_learning/`

Problème : plus l'arbre est profond, plus il est difficile à interpréter

Régression linéaire

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Contribution d'une caractéristique à une prédiction :

$$\beta_k = \frac{y - \sum_{i=1, i \neq k}^{i=n} \beta_i x_i}{x_k}$$

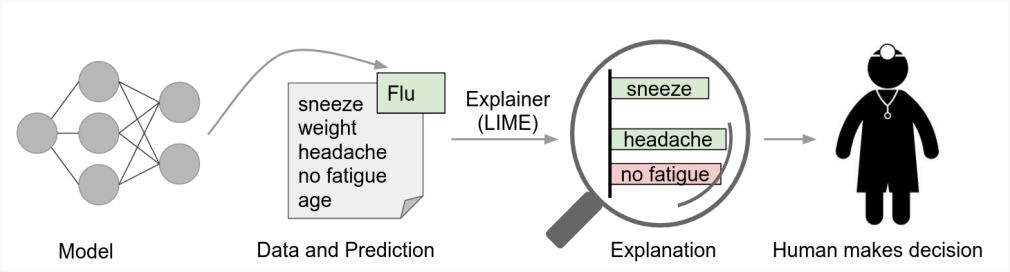
Notations

1. des échantillons $x \in \mathcal{X}$ un espace d'entrée avec la i^{me} caractéristique notée x_i
2. une catégorisation $y \in \mathcal{Y}$ un espace de sortie, y_i itou
3. un programme $f : \mathcal{X} \mapsto \mathcal{Y}$ appris sur un sous-ensemble de $\mathcal{X}$
   - ce programme peut parfois être décomposé en une partie g et une partie h de sorte que $f = h \circ g$
   - on abusera un peu la notation en notant $h(x)$ (dans un réseau de neurones, représente les cartes d'activation à une couche intermédiaire)
4. on note $\nabla_x y$ le gradient de y selon x

Post-hoc

Attribution de caractéristiques

Principe : assigner une notion d'importance à des caractéristiques et indiquer lesquelles ont le plus contribué à la décision



Issu de [RSG16]

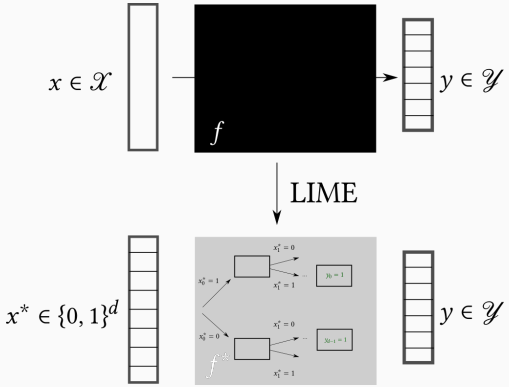
LIME

Local Interpretable Model-agnostic Explanations (LIME) [RSG16] :

idée générale :

1. approche *causale* : modifier les caractéristiques de l'entrée pour estimer leur impact sur la décision finale
   - si non(sneeze) => non(flu), alors sneeze est une donnée importante
2. une fois ces données identifiées, apprendre un programme plus facilement interprétable que le programme source

LIME - cont.

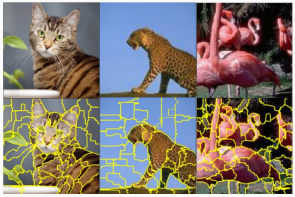


Principe de LIME. La définition de x^* dépend du domaine d'application.

Il doit exister une correspondance $\mathscr{H}\{0,1\}^d \mapsto \mathcal{X}$.

LIME - cont.

Exemples des images : x^* décrit la présence de *d superpixels* sur x. $\mathscr{H}(x_0^* = 0)$ correspond à x avec le superpixel 0 occulté



LIME - principe

1. Prendre un échantillon x
2. Le perturber en ajoutant ou retirant des caractéristiques, créant ainsi un ensemble $\{x^* \in \{0,1\}^d\}$ (qu'on note $\mathcal{X}^*$)
3. Créer un dataset constitué de $\mathcal{X}^*$ et de $f(\mathcal{H}(\mathcal{X}^*))$
4. Pondérer les permutations x^* en fonction de la distance à x
5. Apprendre un modèle interprétable sur ce dataset

L'explication est *fidèle localement* : elle n'explique qu'une prédiction donnée

LIME - avantages

Fonctionne sur tous types de données et de modèles : pas besoin d'avoir accès au programme

LIME - inconvénients

Difficile de définir un voisinage dans l'espace des images

Valeurs de Shapley et théorie des jeux

Il est possible de voir une prédiction sous un autre jour : la théorie des jeux

Valeurs de Shapley et théorie des jeux

Il est possible de voir une prédiction sous un autre jour : la théorie des jeux

Chaque prédiction est un jeu entre chaque caractéristique : de combien contribuent-elles à la prédiction ?

Valeurs de Shapley et théorie des jeux

Il est possible de voir une prédiction sous un autre jour : la théorie des jeux

Chaque prédiction est un jeu entre chaque caractéristique : de combien contribuent-elles à la prédiction ?

$g(x^*)$ est la « contribution totale »

$$g(x^*) = \phi_0 + \sum_{k=1}^{d} \phi_k x_k^*$$

ϕ_k : valeur de Shapley

Valeurs de Shapley - idée générale

Intuition

Évaluer la contribution d'une caractéristique pour une prédiction donnée *par rapport à la moyenne des prédictions*

Valeurs de Shapley - coalitions

Soit $x^* \in \{0,1\}^d$ ($\mathscr{H}(x^*) = x$ ssi $x_i^* = 1 \forall i \in (1..d)$)

Valeurs de Shapley - coalitions

Soit $x^* \in \{0,1\}^d$ ($\mathscr{H}(x^*) = x$ ssi $x_i^* = 1 \forall i \in (1..d)$)

On note l'ensemble $\{x_k | k = 0 \forall k \in (1..d)\}$ une *coalition* (un rassemblement de caractéristiques qui coopèrent pour obtenir une prédiction)

Valeurs de Shapley - garanties théoriques

1. *efficacité* : $\sum_{k=1}^{d} \phi_j = f(x) - \mathrm{E}_{\mathcal{X}}[f(\mathcal{X})]$ (on peut calculer la contribution de chaque caractéristique)
2. *symétrie* : si deux caractéristiques i et j contribuent identiquement à chaque coalition, alors leur valeur de Shapley est égale
3. *nullité* : une caractéristique qui n'influe pas sur la décision a une valeur de Shapley nulle
4. *additivité* : pour un jeu combinant plusieurs prédiction, les valeurs de Shapley s'ajoutent

SHAP

SHAP [LL17] introduit une technique d'échantillonnage plus efficace des valeurs de Shapley, et y lie le formalisme de LIME

SHAP - Principe

1. Prendre un échantillon x
2. Le perturber en ajoutant ou retirant des caractéristiques, créant ainsi un ensemble $\{x^* \in \{0,1\}^d\}$ (qu'on note $\mathcal{X}^*$)
    - Différence par rapport à LIME : on remplace les caractéristiques vides par la valeur moyenne sur tout le dataset
3. Créer un dataset constitué de $\mathcal{X}^*$ et de $f(\mathcal{H}(\mathcal{X}^*))$
4. Pondérer les permutations x^* en fonction de la taille de la coalition (il est plus facile d'estimer l'importance d'une caractéristique quand elle est isolée)
5. Échantillonner de manière répétée

Shapley - avantages

1. théoriquement très solide (toutes les propriétés de valeurs de Shapley)
2. cohérence : si un modèle change de sorte à augmenter la contribution marginale d'une caractéristique, la valeur de Shapley de cette caractéristique augmentera
3. autorise l'analyse de modèles d'ensembles

Shapley - inconvénients

1. lourd en calcul
2. besoin des données d'entrée
3. échantillonnage par permutation de caractéristiques : il y a un risque d'être biaisé vers des exemples irréalistes

RISE

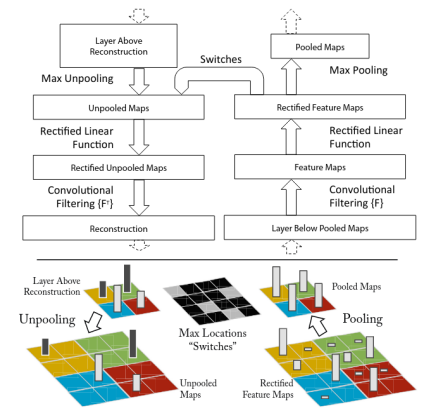
Randomized Input Sampling for Explanation (RISE) [PDS18] occulte des fractions de l'image avant une prédiction, et combine les masques ainsi générés pondéré avec la prédiction de probabilité

$$\frac{1}{\mathbb{E}[\mathrm{M}]} \sum_{m} f(x \cdot m) \cdot m \cdot \mathrm{P}[\mathrm{M} = m]$$

génération initiale de petits masques puis augmentation progressive de la surface pour rester faisable techniquement

Autres approches d'attribution

Proposée initialement dans [ZF14] : technique de « déconvolution » pour reconstruire dans l'espace visuel les caractéristiques apprises par un réseau



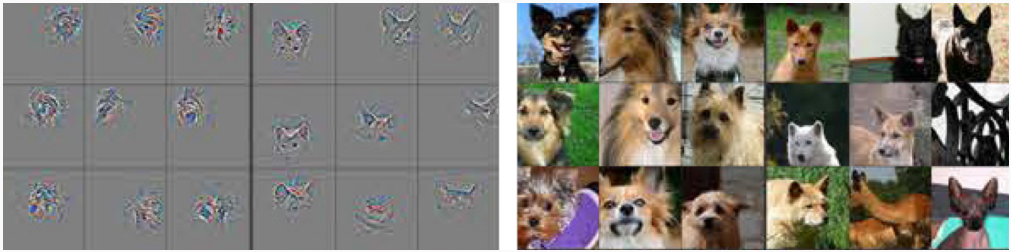
Attribution de concepts

ConceptShap [Yeh+19] proposent un score de « suffisance d'attribution de concepts » et une technique de recherche de concepts (extension de LIME)

Attribution de caractéristiques - bilan

Ne nécessitent pas un accès direct au modèle (seulement entrée et sortie), questions de distance et de voisinages cruciales (et non réglées)

Cartes de chaleur



Source [ZF14]

idée originelle : calculer $\nabla_x y$ et visualiser sur l'espace d'entrée les pixels les plus importants

GRADCAM

GRADCAM [Sel+16] calcule le gradient d'une carte d'activation par rapport à une classe donnée : $\nabla_{h(x)} y_i$, puis les points résultants sur suréchantillonés dans l'espace des entrées

Amélioration GRADCAM++ [Cha+18] qui propose une backpropagation plus complexe

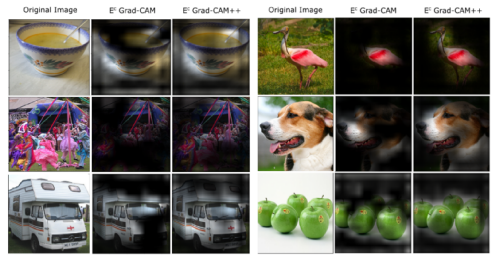


Figure 3 – Issu de [Cha+18]

SMOOTHGRAD

SMOOTHGRAD [Smi+17] $\nabla_{x^*} y$ où on échantillonne x^* est un voisinage gaussien de x

Idée : pour « lisser » les gradients qui varient beaucoup (du fait des évolutions discontinues causées par les ReLU)

Gradients intégrés

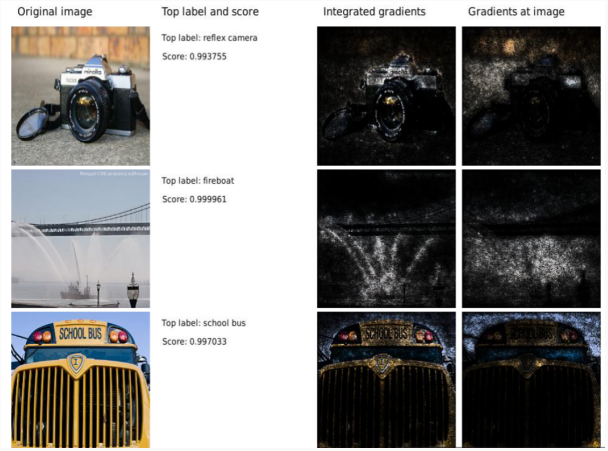
Gradient sur les points du segment reliant x et une image baseline $x^{'}$ [STY17]

$$\mathrm{IG}_i = (x_i - x_i^{'}) \int_{\alpha=0}^{1} \nabla_{x_i} f(x^{'} + \alpha(x - x^{'}))d\alpha$$

s'approxime par Riemann

$$\mathrm{IG}_i \approx (x_i - x_i^{'}) \sum_{k=0}^{m} \nabla_{x_i} f(x^{'} + \frac{m}{k}(x - x^{'})) * \frac{1}{m}$$

Gradients intégrés



Gradients intégrés - suite

plusieurs propriétés théoriques intéressantes

- attribution correcte de caractéristiques pertinentes
- si deux modèles ont la même sortie, les gradients intégrés donneront une explication similaire
- préservation de la symétrie

mais instables en fonction du choix de $x^{'}$

Cartes de chaleur - bilan

1. nécessitent (souvent) de rétropropager un gradient
2. proposent une représentation visuelle
3. mais il y a des défauts (spoiler)

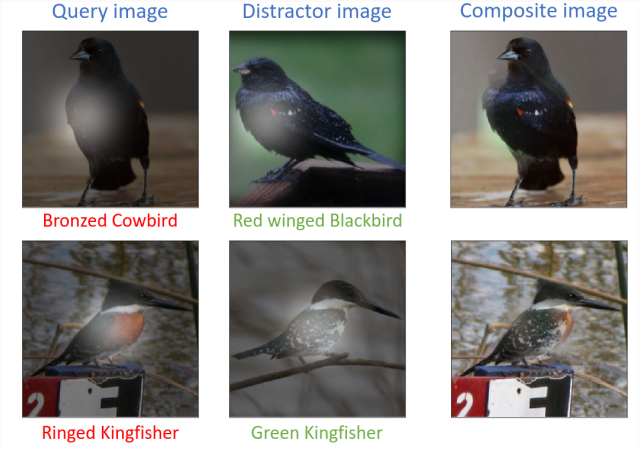
Explication contrastives pour les images

Par exemple [Goy+19]

Trouver le chemin minimum qui amène à un changement de prédiction. Soit x tel que $f(x) = c$, $x^{'}$ tel que $f(x^{'}) = c^{'}$, trouver une image x^{*} telle que $f(x^{*}) = c^{'}$

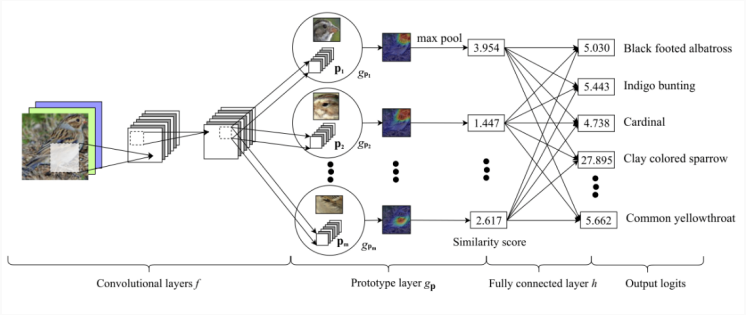
Algorithme glouton de substitution de caractéristiques dans l'espace latent jusqu'à obtenir un changement

Explications contrastives pour les images



Programmes interprétables par construction

Approches par prototypes - ProtoPnet

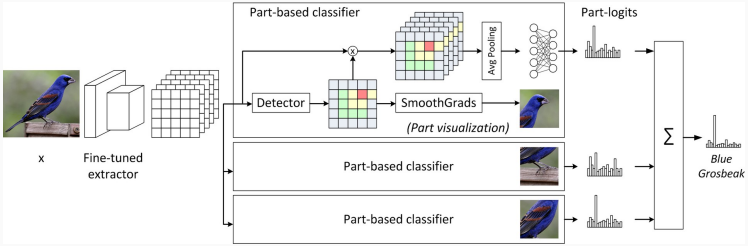


Issu de [Che+19]

Approches par prototypes - ProtoPnet

1. Apprend des « prototypes » et émet une prédiction en fonction de la similarité entre des parties de l'image et le prototype
2. Prototypes appris en étant « bien séparés » (une image a un patch proche d'un prototype de sa classe, tout en s'éloignant de prototypes pas de sa classe)
3. Représentation latente ensuite projetée dans l'espace des images (et ça cause de soucis)

Classification par classe [Xu-+22]



Et plus encore...

1. explications de modèles transformers (travaux en cours dans mon laboratoire)
2. équations différentielles neuronales et modèles à diffusion [Aug+22]

Évaluation et limitations

Comment évaluer ?

Quelques critères regroupés dans [Bod+21]

1. études sur utilisateurs avec protocole expérimental (avec des vrais gens !)
2. invariance des explications pour une même entrée
3. une petite variation dans les entrées doit entraîner une petite variation dans l'explication

Retirer les pixels surlignés par une technique d'interprétabilité et étudier l'évolution de la performance du programme

Comment évaluer ?

Quelques critères proposés dans [Nau+22] (Co12)

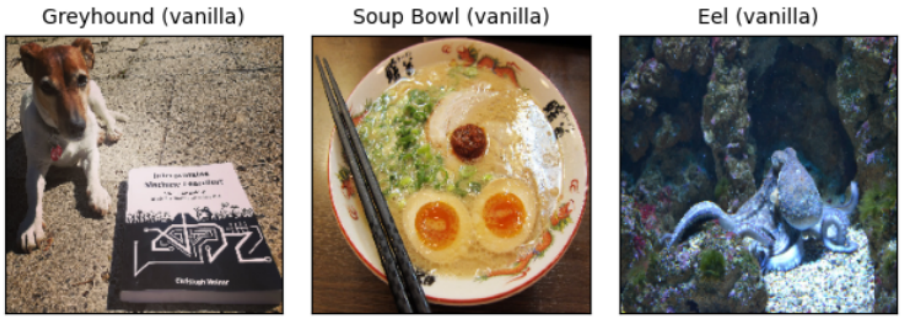
1. *correction* (la plupart des programme *by design*)
2. *cohérence* (invariante à l'implémentation, LIME, Shapley, gradients intégrés)
3. *constrastivité* (méthodes contrefactuelles)
4. *compactness* (LIME avec réduction de taille de feature, la plupart des méthodes de cartes de chaleur)
5. *composabilité* (Shapley)
6. *controllability* (aucune présentée ici)

Évaluation - conclusion

il n'existe pas de métrique absolue :

1. la littérature n'a pas encore convergé vers une évaluation commune des métriques sans humains
2. les études avec utilisateur offrent une idée réelle de l'acceptabilité de la méthode, mais a ses propres soucis

Quelles limitations ?



RE 10.9: Images of a dog classified as greyhound, a ramen soup classified as soup bowl, and an octopus classified as eel.

Issu de [Mol22]

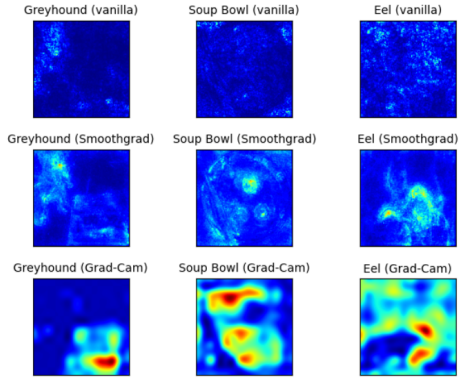


FIGURE 10.10: Pixel attributions or saliency maps for the Vanilla Gradient method, SmoothGrad and Grad-CAM.

Issu de [Mol22]

Le réseau a vraisemblablement pris une mauvaise décision… mais pourquoi ?

Cette explication n'aide pas à ajuster notre modèle mental du fonctionnement du programme (mis à part « ça marche pas »)

Nuance

Extraire une chaîne de causalité et la présenter à une personne constitue une attribution de causalité, mais pas nécessairement une explication [Mil19]

La plupart des méthodes présentées sont une forme d'attribution de causalité, ce qui est souvent insuffisant pour « combler les trous »

« *Comment* la décision a été prise » et « *Pourquoi* la décision a été prise » sont deux questions différentes

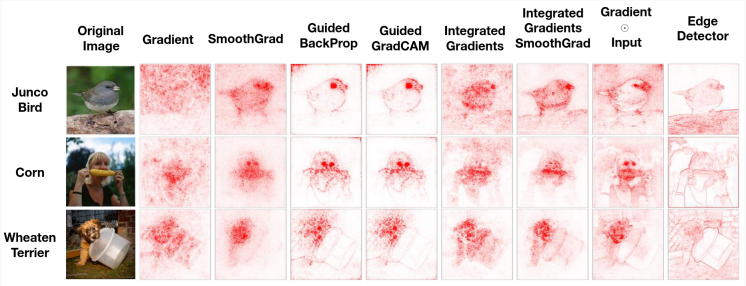
Un regard extérieur

Les concepteurs de méthodes (= des gens comme vous et moi) gagneraient à s'intéresser au corpus des sciences sociales.

Ainsi, [STY17] utilise une approche issue de l'économie, et on voit de plus en plus d'études humaines [1]

1. qui recourent souvent à Amazon Mechanical Turk, une plateforme de microtravail qui créé une précarité difficilement justifiable pour ses travailleurs [Tub21]

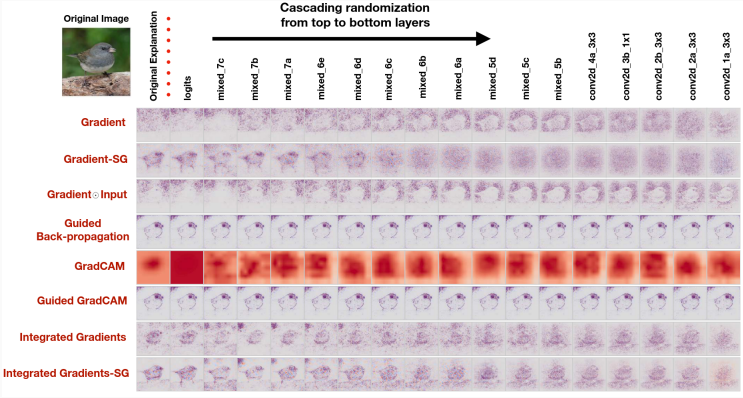
(In)dépendance à la donnée



Issu de [Tom+19]

Certaines techniques produisent un résultat qui varie peu en fonction de la donnée...

(In)dépendance au modèle



Issu de [Tom+19]. Plus on va à droite, plus le réseau comporte de couches aléatoires.

... et en fonction du programme

Voir ce qu'on veut bien voir

Bias de confirmation (Wikitionnaire)

Erreur logique, tendance naturelle de l'individu, consistant à ne retenir que les expériences qui vont dans le sens de ses convictions, à chercher les éléments qui confirment son idée, au lieu de chercher aussi ce qui l'infirme

Une « jolie » carte de chaleur conformera notre idée que le programme fonctionne comme attendu, alors qu'elle ne traduit pas forcément les principes du programmes

Manipulations d'explications [Dom+19]

Une entité pourrait manipuler un réseau pour tromper les méthodes d'explication, sans aucune manière de détecter l'intervention

Cas d'usages : tromper une entité légale, masquer des biais indésirables

Bilan et vers où aller ?

- le besoin d'interprétabilité est clairement établi
- la communauté manque de pratiques communes
- un effort encourageant vers d'autres disciplines (explications contrastives), mais il y a encore beaucoup de chemin
- le caractère discursif de l'explication, combinée avec un modèle génératif de langue, pourrait être très prometteur

On recrute !

Site web : `https://caisar-platform.github.io/website/`
Logiciel libre (LGPLv2) : `https://git.frama-c.com/pub/caisar`
Rapport technique : `https://hal.science/hal-03687211`
Offres :
`https://caisar-platform.github.io/website/positions`
Stages, post-docs, CDD à pourvoir ! julien.girard2@cea.fr

Références

[Aug+22] Maximilian Augustin, Valentyn Boreiko, Francesco Croce et Matthias Hein. *Diffusion Visual Counterfactual Explanations*. 2022. doi : 10.48550/ARXIV.2210.11841. url : https://arxiv.org/abs/2210.11841 (cf. p. 65).

[Bod+21] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi et Salvatore Rinzivillo. « Benchmarking and Survey of Explanation Methods for Black Box Models ». In : *ArXiv* abs/2102.13076 (2021) (cf. p. 67).

Bibliography ii

[Cha+18] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader et Vineeth N Balasubramanian. « Grad-CAM++ : Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks ». In : *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mars 2018. DOI : 10.1109/wacv.2018.00097. URL : https://doi.org/10.1109%2Fwacv.2018.00097 (cf. p. 53).

Bibliography iii

[Che+19] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su et Cynthia Rudin. « *This* Looks like *That* : Deep Learning for Interpretable Image Recognition ». In : *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (2019), p. 8930-8941 (cf. p. 62).

[Dom+19] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller et Pan Kessel. « Explanations Can Be Manipulated and Geometry Is to Blame ». In : *Advances in Neural Information Processing Systems*. T. 32. Curran Associates, Inc., 2019 (cf. p. 79).

Bibliography iv

[Goy+19] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh et Stefan Lee. *Counterfactual Visual Explanations*. 2019. DOI : 10.48550/ARXIV.1904.07451. URL : https://arxiv.org/abs/1904.07451 (cf. p. 59).

[LL17] Scott M. Lundberg et Su-In Lee. « A Unified Approach to Interpreting Model Predictions ». In : *NIPS*. 2017 (cf. p. 44).

[Mil19] Tim Miller. « Explanation in Artificial Intelligence : Insights from the Social Sciences ». In : *Artificial Intelligence* 267 (2019), p. 1-38 (cf. p. 12, 13, 74).

Bibliography v

[Mol22] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2e éd. 2022. URL : https://christophm.github.io/interpretable-ml-book (cf. p. 7, 71, 72).

[Nau+22] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen et Christin Seifert. « From Anecdotal Evidence to Quantitative Evaluation Methods : A Systematic Review on Evaluating Explainable AI ». In : *CoRR* abs/2201.08164 (2022). arXiv : 2201.08164. URL : https://arxiv.org/abs/2201.08164 (cf. p. 7, 9, 68).

Bibliography vi

[PDS18] Vitali Petsiuk, Abir Das et Kate Saenko. « RISE : Randomized Input Sampling for Explanation of Black-box Models ». In : *BMVC*. 2018 (cf. p. 48).

[RSG16] Marco Tulio Ribeiro, Sameer Singh et Carlos Guestrin. « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier ». In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016) (cf. p. 30, 31).

[Sel+16] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh et Dhruv Batra. « Grad-CAM : Why did you say that ? » In : *ArXiv* abs/1611.07450 (2016) (cf. p. 53).

Bibliography vii

[Smi+17] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas et Martin Wattenberg. « SmoothGrad : removing noise by adding noise ». In : *ArXiv* abs/1706.03825 (2017) (cf. p. 54).

[SP17] Andrew D Selbst et Julia Powles. « Meaningful information and the right to explanation ». In : *International Data Privacy Law* 7.4 (déc. 2017), p. 233-242. issn : 2044-3994. doi : 10.1093/idpl/ipx022. eprint : https://academic.oup.com/idpl/article-pdf/7/4/233/22923065/ipx022.pdf. url : https://doi.org/10.1093/idpl/ipx022 (cf. p. 16).

Bibliography viii

[STY17] Mukund Sundararajan, Ankur Taly et Qiqi Yan. « Axiomatic Attribution for Deep Networks ». In : *Proceedings of the 34th International Conference on Machine Learning*. Sous la dir. de Doina Precup et Yee Whye Teh. T. 70. Proceedings of Machine Learning Research. PMLR, juin 2017, p. 3319-3328. URL : https://proceedings.mlr.press/v70/sundararajan17a.html (cf. p. 55, 75).

[Tom+19] Richard J. Tomsett, Daniel Harborne, Supriyo Chakraborty, Prudhvi K. Gurram et Alun David Preece. « Sanity Checks for Saliency Metrics ». In : *ArXiv* abs/1912.01451 (2019) (cf. p. 76, 77).

Bibliography ix

[Tub21] Paola Tubaro. « Disembedded or Deeply Embedded ? A Multi-Level Network Analysis of Online Labour Platforms ». In : *Sociology* (31 jan. 2021), p. 003803852098608. ISSN : 0038-0385, 1469-8684. DOI : 10.1177/0038038520986082. URL : http://journals.sagepub.com/doi/10.1177/0038038520986082 (visité le 10/08/2021) (cf. p. 75).

Bibliography x

[Xu-+22] Romain Xu-Darme, Georges Quénot, Zakaria Chihani et Marie-Christine Rousset. « PARTICUL : Part Identification with Confidence measure using Unsupervised Learning ». Accepted at XAIE : 2nd Workshop on Explainable and Ethical AI – ICPR 2022. Juin 2022. URL : https://hal-cea.archives-ouvertes.fr/cea-03703962 (cf. p. 64).

[Yeh+19] Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister et Pradeep Ravikumar. *On Completeness-aware Concept-Based Explanations in Deep Neural Networks*. 2019. DOI : 10.48550/ARXIV.1910.07969. URL : https://arxiv.org/abs/1910.07969 (cf. p. 50).

[ZF14] Matthew D. Zeiler et Rob Fergus. « Visualizing and Understanding Convolutional Networks ». In : *Computer Vision – ECCV 2014*. Sous la dir. de David Fleet, Tomas Pajdla, Bernt Schiele et Tinne Tuytelaars. Cham : Springer International Publishing, 2014, p. 818-833. isbn : 978-3-319-10590-1 (cf. p. 49, 52).