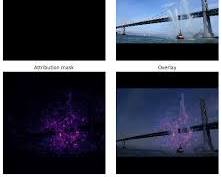
BARD prompt for explainable ML algorithm

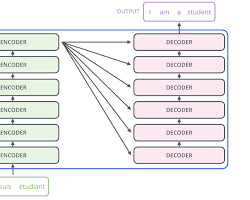
The latest updates of algorithms for interpreting machine learning are still under development, but some of the most promising recent advances include:

* **Integrated gradients:** This method computes the gradient of the model's prediction with respect to the input, which can be used to identify the features that are most important for the prediction.

[Opens in a new window [](https://www.tensorflow.org/tutorials/interpretability/integrated_gradients)www.tensorflow.org](https://www.tensorflow.org/tutorials/interpretability/integrated_gradients)

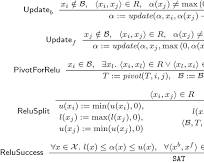
Integrated gradients algorithm illustration

* **DeepSHAP:** This method extends SHAP to deep learning models, by computing the Shapley values for each layer of the model.
* Can't find image. Sorry, Bard is experimental.
* **Explanatory Transformer:** This method uses a transformer model to generate explanations for the predictions of a machine learning model.

[Opens in a new window jalammar.github.io](http://jalammar.github.io/illustrated-transformer/)

Explanatory Transformer algorithm illustration

* **Reluplex:** This method is a formal verification technique that can be used to check whether a machine learning model is making fair predictions.

[Opens in a new window [](https://www.researchgate.net/figure/Additional-derivation-rules-for-the-abstract-Reluplex-algorithm_fig5_352900331)www.researchgate.net](https://www.researchgate.net/figure/Additional-derivation-rules-for-the-abstract-Reluplex-algorithm_fig5_352900331)

Reluplex algorithm illustration

These are just a few of the many recent advances in machine learning interpretability. As research in this area continues, we can expect to see even more powerful and effective methods for understanding and explaining the inner workings of machine learning models.

It is important to note that there is no single "best" algorithm for interpreting machine learning. The best algorithm for a particular application will depend on the specific model, the data, and the desired level of interpretability. However, the methods mentioned above are a good starting point for exploring the possibilities of machine learning interpretability.

<https://link.springer.com/article/10.1007/s10618-023-00933-9>

Bodria, F., Giannotti, F., Guidotti, R. *et al.* Benchmarking and survey of explanation methods for black box models. *Data Min Knowl Disc* **37**, 1719–1778 (2023). <https://doi.org/10.1007/s10618-023-00933-9>

L'essor des modèles sophistiqués d'apprentissage automatique en boîte noire dans les systèmes d'intelligence artificielle a suscité le besoin de méthodes d'explication qui révèlent le fonctionnement de ces modèles de manière compréhensible pour les utilisateurs et les décideurs. Sans surprise, l'état de l'art présente actuellement une pléthore d'explicateurs fournissant de nombreux types d'explications différents. Dans le but de fournir une boussole aux chercheurs et aux praticiens, cet article propose une catégorisation des méthodes d'explication du point de vue du type d'explication qu'elles fournissent, en tenant également compte des différents formats de données d'entrée. L'article prend en compte les explicateurs les plus représentatifs à ce jour, et examine également les similitudes et les divergences des explications renvoyées en fonction de leur aspect visuel. Un site web accompagnant l'article est fourni pour une mise à jour continue des nouveaux explicateurs au fur et à mesure de leur apparition. En outre, un sous-ensemble des explicateurs les plus robustes et les plus largement adoptés est comparé à un répertoire de mesures quantitatives.

Il existe un risque inhérent que le fait de s'appuyer sur des modèles opaques conduise à adopter des décisions que nous ne comprenons pas parfaitement ou, pire encore, à violer des principes éthiques ou des normes juridiques. Ces risques sont particulièrement importants dans les scénarios de prise de décision à fort enjeu, tels que la médecine, la justice, la finance, le recrutement, l'accès aux prestations publiques, etc. (Rudin 2019) ; le manque de transparence et de confiance peut expliquer le taux d'adoption relativement faible des systèmes actuels d'aide à la décision basés sur l'IA dans les domaines mentionnés. En outre, les entreprises qui intègrent des modèles ML à boîte noire dans leurs produits et applications d'IA risquent de subir une perte potentielle de sécurité et de confiance (Chouldechova 2017).

De plus, afin de fournir au lecteur un examen complet des méthodes d'IAO, nous complétons cette section en présentant les définitions de l'explication, de l'interprétabilité, de la transparence et de la complexité qui figurent dans la documentation :

L'explication (Arrieta et al. 2020 ; Guidotti et al. 2019c) est une interface entre les humains et un décideur IA qui est à la fois compréhensible pour les humains et une approximation précise de l'IA. Par conséquent, l'explicabilité est la capacité à fournir une explication valable.

L'interprétabilité (Guidotti et al. 2019c), ou la compréhensibilité (Gleicher 2016), est la capacité des parties prenantes à comprendre les aspects pertinents du processus de modélisation. L'interprétabilité et la compréhensibilité sont liées à l'évaluation de la complexité du modèle.

La transparence (Arrieta et al. 2020), ou de manière équivalente la compréhensibilité ou l'intelligibilité, est la capacité d'un modèle à être lui-même interprétable. Ainsi, le modèle permet à un humain de comprendre directement son mécanisme interne et son processus de décision.

La complexité (Doshi-Velez et Kim 2017) est le degré d'effort requis par un utilisateur pour comprendre une explication. La complexité peut prendre en compte les antécédents de l'utilisateur ou les éventuelles limitations de temps nécessaires à la compréhension.

# #### different type of models

Explicatifs pour les boîtes noires classant les données d'images, triés par type d'explication : Cartes d’attention (saliency map, SM), Attributions de concepts (CA), Contrefactuels (CF) et Prototypes (PR).

Pour chaque méthode, il est indiqué s'il est possible de l'utiliser pour des images (IMG) uniquement ou pour TOUT type de données, s'il s'agit d'un modèle intrinsèque (IN) ou post-hoc (PH), local (L) ou global (G), et s'il s'agit d'un modèle agnostique (A) ou d'un modèle spécifique (S).

Cartes de saillance (SM)

C’est une image dans laquelle la luminosité d'un pixel représente le degré de saillance de ce pixel. Formellement, une carte de saillance est modélisée comme une matrice S, dont les dimensions sont les tailles de l'image pour laquelle nous voulons expliquer une décision, et les valeurs 𝑠𝑖𝑗 sont les valeurs de saillance des pixels ij. Plus la valeur de 𝑠𝑖𝑗 est grande, plus la saillance de ce pixel est importante.

LIME : Ribeiro et al. (2016), Tout type de données, Post-hoc, Local, Agnostique

Lime est un explicateur local post-hoc agnostique de modèle, il renvoie des saliency map (SM) basés sur la segmentation où chaque segment est appelé superpixel. Après la segmentation de l'image d'entrée, lime adopte une représentation vectorielle à un coup de l'image où l'image d'entrée est un vecteur composé de m uns si m est le nombre de segments identifiés. Ensuite, il crée le voisinage synthétique en remplaçant aléatoirement les superpixels par une couleur uniforme, éventuellement neutre, et en stockant également les représentations à un coup. Le voisinage des images synthétiques est ensuite introduit dans la boîte noire, et la représentation à un instant du voisinage, ainsi que la prédiction de la boîte noire, sont utilisées pour former un modèle linéaire clairsemé. Enfin, les coefficients du modèle linéaire sont utilisés pour déterminer l'importance des superpixels. Des exemples d'explications renvoyées par lime sont présentés dans la deuxième ligne de la figure 7. Un aspect critique pour obtenir une bonne explication avec cette approche est le choix de l'algorithme de segmentation et de ses hyperparamètres. En effet, pour les images à faible résolution, la segmentation de lime ne fonctionne pas de manière standard, ce qui conduit l'algorithme à sélectionner toutes les images comme un superpixel. Récemment, de nombreuses recherches ont amélioré et étendu lime pour les classificateurs d'imagesNote de bas de page 18 (Shi et al. 2020 ; Peltola 2018 ; Zafar et Khan 2019 ; Bramhall et al. 2020).

Grad-cam, Selvaraju et al. (2020), Image uniquement, Post-hoc, Local, Spécifique

GRAD-CAM est un explicateur local post-hoc spécifique au modèle pour les données d'image. Il utilise les informations de gradient qui circulent dans la dernière couche convolutive d'un CNN pour attribuer des valeurs de saillance à chaque neurone pour une décision particulière. Les couches convolutives retiennent naturellement les informations spatiales dans les couches entièrement connectées, de sorte que l'on suppose que les dernières couches convolutives présentent le meilleur compromis entre la sémantique de haut niveau et les informations spatiales détaillées. Pour créer le SM, grad-cam prend les cartes de caractéristiques créées à la dernière couche du réseau convolutif a. Ensuite, il calcule le gradient d'une sortie d'une classe particulière 𝑦𝑐 pour chaque activation de carte de caractéristiques k, c'est-à-dire ∂𝑦𝑐/∂𝑎𝑘. Cette équation renvoie un tenseur de dimensions [k, v, u] où k est le nombre de cartes de caractéristiques et u et v sont la hauteur et la largeur de l'image. grad-cam calcule la valeur de saillance pour chaque carte de caractéristiques en regroupant les dimensions de l'image. La carte thermique finale est calculée comme une somme pondérée de ces valeurs. Il en résulte une carte thermique grossière de la même taille que les cartes de caractéristiques convolutives. Une technique d'échantillonnage ascendant est appliquée au résultat final pour produire une carte de la dimension initiale de l'image. La figure 7 montre clairement que ce style de carte thermique à gros grain est très caractéristique de la grad-cam. Ces cartes thermiques mettent en évidence des parties très différentes de l'image par rapport aux autres méthodes.

## Attribution de concepts

La plupart des modèles ML sont conçus pour fonctionner sur des caractéristiques de bas niveau comme les bords et les lignes d'une image qui ne correspondent pas à des concepts de haut niveau que les humains peuvent facilement comprendre. Dans (Adebayo et al. 2018 ; Yang et Kim 2019), les auteurs ont souligné que les explications basées sur les caractéristiques appliquées aux modèles de boîte noire de pointe peuvent produire des explications non sensibles. Par exemple, nous pouvons considérer les SM comme des explications de bas niveau pour les images, car ils attribuent à chaque pixel une valeur de saillance. Bien qu'il soit possible de regarder chaque pixel et de déduire sa valeur numérique, celle-ci n'a que peu ou pas de sens pour les humains : nous ne disons pas que le 5e pixel d'une image a une valeur de 28. En revanche, les méthodes d'attribution de concepts (AC) quantifient, par exemple, la contribution du concept "rayures" à la prédiction de la classe "zèbre". En effet, les méthodes d'explication basées sur l'AC construisent l'explication sur la base de concepts définis par l'homme plutôt que de représenter les entrées sur la base de caractéristiques et d'états du modèle interne (activation). Par conséquent, cette idée de caractéristiques de haut niveau peut être plus familière aux humains, qui peuvent être plus susceptibles de l'accepter (Hartmann et al. 2022 ; Renard et al. 2019). Formellement, étant donné un ensemble d'images appartenant à un concept [𝑥(1),𝑥(2),...,𝑥(𝑖)] avec 𝑥(𝑖)∈𝐶

Les méthodes de CA peuvent être considérées comme une fonction 𝑓 :(𝑏,[𝑥(𝑖)])→𝑒 qui attribue un score e au concept C sur la base des prédictions et des valeurs de la boîte noire b sur l'ensemble [𝑥(𝑖)].

TCAV, Testing with Concept Activation Vectors (Kim et al. 2018) est un explicateur global post-hoc agnostique du modèle pour les classificateurs d'images qui fournit une explication quantitative de l'importance d'un concept pour la prédiction. Dans tcav, chaque concept est représenté par un vecteur particulier appelé Concept Activation Vectors (CAVs) créé par l'interprétation d'un état interne d'un réseau neuronal en termes de concepts adaptés aux humains. tcav utilise des dérivés directionnels pour quantifier le degré auquel un concept défini par l'utilisateur est vital pour un résultat de classification. Par exemple, la sensibilité d'une prédiction de "zèbre" à la présence de "rayures". tcav nécessite deux ingrédients principaux : (i) des entrées contenant des concepts et des échantillons négatifs, c'est-à-dire des entrées aléatoires, et (ii) des modèles ML pré-entraînés sur lesquels les concepts sont testés. Les entrées contenant des concepts et les entrées aléatoires sont introduites dans la boîte noire pour obtenir des prédictions qui seront utilisées par tcav pour tester la capacité du modèle d'apprentissage automatique à capturer un concept particulier. Ensuite, un classificateur linéaire est entraîné pour distinguer l'activation du réseau due à des entrées contenant des concepts de celle due à des entrées aléatoires. Le résultat de cette formation est un vecteur d'activation de concept (VAC). Une fois les VAC définis, la dérivée directionnelle de la probabilité de classe le long des VAC peut être calculée pour chaque instance appartenant à une classe. L'"importance du concept" pour une classe est calculée comme une fraction des instances de la classe qui sont activées positivement par les entrées contenant le concept par rapport aux entrées aléatoires.

Plusieurs extensions de tcav sont présentes dans la littérature. ACE, Automated Concept-based Explanation (Ghorbani et al. 2019), est l'évolution de tcav, et il ne nécessite pas de concepts en entrée puisqu'il peut les découvrir automatiquement. Un ensemble d'images de la même classe est segmenté avec plusieurs résolutions résultant en un pool de segments provenant tous de la même classe. L'espace d'activation d'une couche de goulot d'étranglement d'un classificateur CNN est ensuite utilisé comme espace de similarité. Ensuite, les segments similaires sont regroupés dans l'espace d'activation et, pour chaque concept, son score d'importance tcav est calculé en fonction de ses exemples de segments.

ConceptSHAP (Yeh et al. 2020) est une évolution de SHAP qui tente de définir un score d'importance pour chaque concept découvert. conceptshap trouve l'importance de chaque concept individuel à partir d'un ensemble de m vecteurs de concepts 𝐶𝑠={𝑐1,𝑐2,...,𝑐𝑚} en utilisant les valeurs de Shapley. À l'instar de l'ace, le concepthap vise à regrouper les concepts dans certaines régions spatiales cohérentes. CaCE, Causal Concept Effect (Goyal et al. 2019) est une autre variante de tcav qui examine l'effet causal de la présence ou de l'absence de concepts de haut niveau sur la prédiction en boîte noire. En effet, tcav peut souffrir de la confusion des concepts qui pourrait se produire si les instances de données d'apprentissage ont plusieurs classes, même avec une faible corrélation. PACE (Kamakshi et al. 2021) est une variante d'ace qui introduit l'utilisation de la boîte noire pour identifier les concepts en extrayant automatiquement de petites sous-régions de l'image, appelées concepts, pertinentes pour la prédiction de la boîte noire.

icnn, Shen et al. (2021), IMG, IN, G, S

ICNN, an interpretable method to modify traditional CNNs into *interpretable CNN*. In an interpretable CNN, each filter in a high convolutional layer represents a specific object part. An interpretable CNNs use the same training data as ordinary CNNs without additional annotations of object parts or textures for supervision. The interpretable CNN automatically assigns each filter in a high convolutional layer with an object part during the learning process. The explicit knowledge of interpretable CNN can help people understand their logic, i.e., what patterns are memorized by the CNN for prediction.

Les explications basées sur l'attribution de concepts sont le type d'explication le plus connu et le plus utilisé. Cependant, elles sont très fragiles et souffrent de problèmes d'interprétation car leur utilisation d'un point de vue cognitif n'est pas claire. D'autre part, l'utilisation de concepts est une nouvelle approche qui cherche à surmonter les problèmes liés aux SM en produisant des explications de haut niveau qui sont plus compréhensibles pour l'utilisateur final. Les explications basées sur des concepts sont un type très récent d'explications pour les images, et elles présentent des améliorations potentielles. Il s'agit d'un premier pas dans la direction d'explications semblables à celles des humains. Les concepts conviviaux permettent d'élaborer des explications simples et utiles. Les humains doivent encore faire correspondre les images aux concepts, mais c'est un petit prix à payer pour améliorer l'interaction homme-machine. En général, les explicateurs d'attribution de concepts calculent un score qui évalue la probabilité qu'un concept sélectionné ait influencé la prédiction. Le principal problème, pour l'instant, est que ce concept doit être sélectionné manuellement. tcav est l'algorithme le plus avancé. Cependant, les concepts doivent être fournis sous la forme d'un ensemble d'images reflétant le concept sélectionné, par exemple un ensemble d'images de glace, un ensemble de personnes et un ensemble de rayures. Plusieurs variantes de tcav ont été introduites pour automatiser la phase de sélection des concepts. PACE est la recherche la plus avancée lorsqu'il s'agit de trouver des concepts concernant le comportement de la boîte noire. Les chercheurs s'efforcent de trouver des concepts de manière automatisée, mais pour l'instant, il n'existe pas d'explicateur permettant de s'assurer qu'un concept est humainement compréhensible.