Ancres vs Attention : comparaison de méthodes d'explicabilité des réseaux profonds sur un cas d'usage réel

Gaëlle Jouis, Harold Mouchère, Fabien Picarougne, Alexandre Hardouin

Ancres vs Attention : comparaison de méthodes d'explicabilité des réseaux profonds sur un cas d'usage réel

Gaëlle Jouis[*,**] , Harold Mouchère[*],
Fabien Picarougne[*], Alexandre Hardouin[**]

[*]Université de Nantes, LS2N, Polytech Nantes, Rue Christian Pauc, 44300 Nantes
harold.mouchere, fabien.picarougne@ls2n.fr
[**]Pôle Emploi, Direction des Systèmes d'Information, 2 rue Konrad Adenauer, 44200 Nantes
gaelle.jouis, alexandre.hardouin@pole-emploi.fr

Résumé. Les progrès récents de l'intelligence artificielle explicable (XAI) ont conduit à l'arrivée de nombreuses méthodes différentes afin d'améliorer l'explicabilité des algorithmes d'apprentissage profond. Avec de nombreuses options disponibles, et la nécessité d'adapter celles existantes à de nouveaux problèmes, il peut être difficile choisir la bonne méthode pour générer des explications. Cet article présente une approche objective pour comparer deux méthodes d'explicabilité différentes. Ces méthodes sont appliquées à un cas d'usage de la littérature et une application dans une administration française, Pôle Emploi.

1 Introduction

L'intelligence artificielle est en plein essor depuis quelques années. L'apprentissage profond, en particulier, a prouvé son efficacité pour de nombreuses tâches, telles que le traitement d'image, la reconnaissance d'objets et le traitement du langage naturel Zou et al. (2019). Contrairement à d'autres approches comme les modèles linéaires, les modèles d'apprentissage profond sont considérés comme des *boîtes noires*, car leur fonctionnement est opaque. Ainsi, expliquer le résultat d'un algorithme d'apprentissage profond est une tâche difficile.

Cet article compare deux méthodes d'explication différentes, appliquées à un cas d'usage. Ces méthodes sont les Ancres, une méthode boîte noire, et une méthode dite boîte blanche basée sur le méchanisme d'Attention. Ces deux approches sont populaires et basées sur des mécanismes différents. Le cas d'usage basé sur un besoin réel, LEGO, et un deuxième cas d'usage de la littérature, YELP, sont utilisés dans ce processus de qualification, suggérant des bonnes pratiques pour la comparaison de méthodes d'explicabilité.

Le cas d'usage LEGO répond à un besoin concret de classification de texte pour une institution française, *Pôle Emploi*. le problème adressé est la détection d'offres d'emploi non conformes basé sur de l'apprentissage automatique. L'institution a une obligation légale de transparence sur ses algorithmes et cherche à fournir des explications parallèlement aux résultats de l'outil. Le "pourquoi" et le "comment" de l'intelligence artificielle explicable (XAI) sont des sujets traités par la communauté scientifique. À notre connaissance, les solutions pro-

posées ne sont pas systématiquement évaluées. Ainsi, choisir la méthode qui conviendrait le mieux à un projet d'intelligence artificielle (IA) particulier n'est pas une tâche simple.

2 État de l'art

2.1 Intelligence artificielle explicable

Il est possible de regrouper les nombreuses approches d'explicabilité existantes selon une logique globale proposée dans Gilpin et al. (2018) et Guidotti et al. (2018). Ainsi, trois catégories sont définies :

1. Expliquer un modèle *boîte noire* basé sur ses entrées et ses sorties.
2. Observer les mécanismes internes d'un système (*boîte grise*) après son apprentissage.
3. Concevoir une solution transparente (*boîte blanche*) générant ses propres explications.

L'explication des modèles boîte noire induit l'utilisation d'une approximation via un second modèle, plus interprétable, dont LIME et les Ancres, respectivement présentées dans Ribeiro et al. (2016, 2018) sont parmi les plus connues. *LIME* (Local Interpretable Model-agnostic Explanations) est une approximation d'un modèle boîte noire grâce à une régression linéaire. Le résultat de cette régression propose une pondération des entrées, correspondant l'importance relative des termes. De même, la méthode des Ancres explique un résultat au travers d'une règle. La règle présente un ensemble de mots conduisant à la décision du modèle. Ces méthodes sont conçues pour expliquer une seule instance à la fois et ne sont valides que pour les exemples proches.

Le fonctionnement interne des réseaux de neurones à convolution (CNN) a été analysé dans Zeiler et Fergus (2014). Les auteurs ont utilisé un réseau de neurones qu'ils ont nommé *Deconvolutional Network* pour visualiser des modèles qui activent les neurones en fonction des couches. Également basés sur les CNN, les auteurs de Selvaraju et al. (2017) combinent leurs travaux à ceux de Zeiler et Fergus (2014), pour détecter les régions et les motifs d'une image aidant à la détection de classe. Un travail similaire a été fait sur l'analyse sémantique avec les réseaux LSTM (Long Short-Term Memory), dans Karpathy et al. (2015).

Les solutions transparentes, quant à elles, sont inhérentes au modèle développé. Dans Lin et al. (2017), les auteurs créent un plongement de mots basé sur l'attention, appelé "*Structured self-attentive embedding*". Les associés à des poids d'attention élevés sont les plus utilisés par le modèle pour classer le texte. Une autre visualisation basée sur l'attention est présentée dans Olah et Carter (2016). Par rapport aux stratégies type boîte noire, qui réalisent une approximation du modèle entraîné, le mécanisme d'attention est un élément interne du modèle transparent. Après entrainement, il n'y a pas besoin de calcul supplémentaire car l'inférence génère également les poids d'attention, pouvant être utilisés pour générer des explications.

2.2 Évaluer les explications

Deux grands écoles sont présentes dans la littérature : 1) les approches centrées sur les critères et métriques, 2) les approches centrées sur des utilisateurs.

Critères et métriques Dans la littérature, les explications ou les modèles qui les génèrent, appelés modèles proxy, sont souvent évalués avec des ensembles de critères et métriques. L'un des critères les plus utilisés est la fidélité du modèle proxy au modèle boîte noire, mesurée avec le taux de reconnaissance ou le F1-score selon Guidotti et al. (2018); Ribeiro et al. (2018). La couverture peut être mesurée comme le nombre d'instances qui sont en accord avec une explication comme dans Ribeiro et al. (2018). D'autres métriques peuvent également être utilisées, par exemple la taille du modèle proxy ou encore son nombre de paramètres, comme le relèvent les auteurs de Guidotti et al. (2018). Dans le cas des explications en langage naturel, il est possible d'utiliser un score de lisibilité tel que le score *Flesch-Reading-Ease*, utilisé dans Costa et al. (2018). Lorsque les explications requises sont disponibles, les explications attendues et obtenues peuvent être comparées comme des ensembles de caractéristiques. Le calcul de l'intersection sur l'union (IoU) donne un score de 0 à 1. Une IoU de 1 signifie que les explications sont identiques. Cette métrique est utilisée dans Bau et al. (2017) pour évaluer l'interprétabilité dans le domaine de l'analyse d'images. Si seules quelques explications sont possibles, le cas peut être considéré comme un problème de classification, et les mesures de performances habituelles peuvent être utilisées comme dans Codella et al. (2019).

L'évaluation basée sur des métriques permet de travailler sur de grands ensembles de données de test. En s'affranchissant des utilisateurs, il est également plus rapide et moins coûteux de développer des évaluations quantitatives sur n'importe quelle méthode d'explicabilité. Cependant, les explications sont conçues pour être une interface entre les algorithmes et les humains, elles doivent donc être évaluées par, ou avec des humains.

Tests utilisateurs Une définition de l'explicabilité est la capacité d'un humain à comprendre les décisions d'un système étant donné un contexte particulier Miller (2019). Si l'humain est au coeur de la définition de l'explicabilité, il est pertinent de prendre en compte son l'évaluation plutôt que se limiter aux mesures de la section précédente. Lors de la réalisation d'une étude utilisateur sur les explications du modèle, l'évaluation peut être objective ou subjective. L'évaluation subjective peut être un sondage demandant aux utilisateurs s'ils sont satisfaits d'une explication donnée, ou quelle explication préfèrent-ils parmi quelques-uns Ribeiro et al. (2018). On pourrait aussi leur demander de choisir entre deux classificateurs, l'un étant nettement meilleur que l'autre, compte tenu uniquement de leurs explications Ribeiro et al. (2016). Ces évaluations sont appropriées lorsque le but est d'améliorer l'acceptation d'un modèle. D'un autre côté, des métriques objectives peuvent être extraites des études d'utilisateurs. Dans Iyer et al. (2018), les utilisateurs reçoivent une explication et doivent prédire la prochaine sortie du système. Considérant les réponses des utilisateurs comme des résultats de classificateurs binaires, les auteurs calculent une courbe Roc et sa zone sous la courbe pour mesurer le succès de leurs explications. Lorsque l'utilisateur doit prédire la sortie du modèle, le temps de réponse de l'utilisateur peut être utilisé comme une mesure de la confiance de l'utilisateur Ribeiro et al. (2018).

3 Expériences

Nous voulons ici comparer deux méthodes d'explication : la génération d'Ancres sur n'importe quel modèle de Ribeiro et al. (2018) et l'utilisation de l'attention avec un modèle transparent de Lin et al. (2017). Les méthsodes seront appliquées à deux cas d'usage, LEGO et YELP,

qui seront détaillés dans les sections suivantes. Pour chaque cas d'usage, un modèle transparent basé sur l'attention sera formé et des ancres seront générées sur les prédictions de ce même modèle. En s'inspirant de Ribeiro et al. (2018) pour l'exemple, prenons la phrase "Ce film n'est pas mauvais", qui est classée "positive" par un modèle basé sur l'attention. L'explication des Ancres serait $A = \{pas, mauvais\} \rightarrow Positive$. Chaque mot de la phrase possèderait un poids d'attention, et "pas" et "mauvais" auraient les poids les plus élevés.

La génération d'Ancres se fait avec la bibliothèque python développée par les auteurs de Ribeiro et al. (2018). Suite aux recherches de Lin et al. (2017), un réseau de neurones avec un bi-LSTM et le même mécanisme d'attention a été conçu. L'architecture est décrite dans le tableau ci-dessous (cf. Table 1). L'adaptation du réseau à chaque cas d'usage a entraîné des différences de dimensions, qui sont détaillées dans les 3eme et 4eme colonnes du tableau 1. Le mécanisme d'attention aboutit à une matrice d'attention A, qui est la sortie de la couche 5 (cf. Table 1). Les mots d'intérêt sont filtrés en utilisant un seuil t sur les valeurs d'attention. Pour le cas d'usage LEGO, lorsque le modèle ne prédit aucun rejet, l'explication est considérée vide.

TAB. 1 – *Architecture des réseaux de classification pour YELP et LEGO.*

ID	**Type de couche**	**YELP**	**LEGO**	**Commentaires**
1	Couche d'entrée	300	80	La taille est le nombre de mots dans les textes
2	Plongement de mots	100	300	Plongements de mots, respectivement Word2Vec et GloVe
3	Bi-LSTM	u =150	u =50	La sortie est la matrice d'états cachés H
4	Couche dense 0	d_a =350	d_a =300	Activation $tanh$
5	Couche dense 1	r =1	r =1	La sortie est la matrice d'attention A
6	Attention et moyenne	sortie : $[2u, r]$	sortie : $[2u, r]$	Combinaison de l'attention et de la couche cachée, $M = A^T * H$
7	Couche dense 2	1000		Activation $ReLu$, pour YELP uniquement
8	Couche dense 3	5	28	Couche de sortie

3.1 LEGO

Pôle Emploi est un établissement public à caractère administratif, mettant en relation recruteurs et demandeurs d'emploi. L'un de ses outils vise à rejeter automatiquement les offres d'emploi non conformes. En effet, *Pôle Emploi* est légalement tenu de rejeter les offres non conformes au Code du travail ou discriminatoires. Le jeu de données d'entrainement contient 480000 phrases extraites d'offres réelles. La recherche du motif de rejet est une tâche de classification multiclasses, 28 motifs étant ciblés dans cette étude. Les offres utilisées pour la formation du modèle sont déjà labelisées dans la base de données de *Pôle Emploi*, avec des labels prédit par le système à base de règles déjà en place.

Pour ce classifieur, la matrice de plongement de mots utilisée est un plongement GloVe de 300 dimensions. [1] L'optimiseur Adam est utilisé, avec un taux d'apprentissage de $0,0005$. Ce réseau atteint un taux de reconnaissance de $83,67\%$ sur son ensemble de test.

Le système existant produit des erreurs. Ainsi, pour analyser avec précision les explications, un ensemble de test corrigé était nécessaire. Comme la correction des labels prend du temps, un sous-ensemble de 208 phrases a été labellisé manuellement et associé à l'explication souhaitée, appellée "vérité terrain" dans la suite de cet article. Cette explication consiste à mettre en évidence les mots-clés qui ont conduit au rejet. Étant donné que les explications n'ont de sens que pour les offres non conformes, les explications pour les offres conformes sont considérées comme vides. La distribution de classes du monde réel n'a pas été respectée et les explications conformes sont sous-représentées dans ces données de test. Par conséquent, le taux de reconnaissance du modèle pour ce jeu de données de test est inférieure et non pertinente ($70,67\%$).

3.2 YELP

L'ensemble de données YELP contient des avis d'utilisateurs sur des commerces, associés à des notes allant de 1 à 5 étoiles. Le jeu de données d'apprentissages contient 453 600 avis. Comme le montre le tableau 1, le plongement de mots est un Word2Vec de 100 dimensions basé sur de l'anglais. [2] L'optimiseur Adam est utilisé, avec un taux d'apprentissage de $0,0005$. Ce réseau atteint un taux de reconnaissance de $74,63\%$ sur son ensemble de test. En comparaison, les auteurs de Lin et al. (2017) présentent dans leur article un taux de reconnaissance de $64,21\%$ sur leur propre ensemble de tests. Comme l'explication des ancres sur des textes volumineux conduit souvent à des problèmes de mémoire, les ancres ont été appliquées à un sous-ensemble de 1060 avis sur les 2653 de l'ensemble de test complet.

4 Evaluer les explications

4.1 Analyse quantitative

Pour le cas d'usage LEGO, chaque méthode est comparée à la vérité terrain. Pour obtenir des mesures équitables, les mots vides de sens ne sont pas pris en compte. Comme l'évaluation du modèle n'est pas le but de cette expérimentation, le jeu de test est un sous-ensemble de 147 phrases sur 208, qui ont été correctement prédites. Un seuil t est utilisé pour filtrer les mots. t est déterminé en optimisant l'IoU sur l'ensemble de test, comme le montre la Figure 1. Le graphique indique que les mots avec une attention supérieure ou égale à 0,15 sont une bonne explication dans le cas d'usage LEGO.

Comme vu dans Bau et al. (2017), IoU permettra de comparer la vérité terrain à l'explication générée, avec les Ancres ou l'Attention. Le taux de reconnaissance et le F1-score sont également affichés dans le tableau 2, ainsi que le rappel et la précision utilisés dans le score F1. Le rappel est une métrique intéressante car elle n'est pas affectée par les vrais négatifs. Les explications des Ancres et l'Attention sont également comparées les unes aux autres, vérifiant ainsi qu'elles donnent des résultats similaires. En l'absence de vérité terrain, les métriques

1. https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.fr.300.vec.gz
2. https://wikipedia2vec.github.io/wikipedia2vec/pretrained/

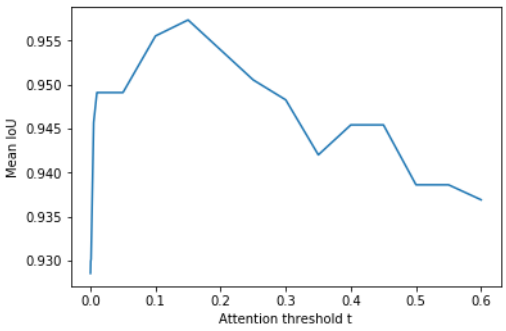

FIG. 1 – *IoU moyenne entre les explications Attention et vérité terrain dans l'ensemble de test, pour faire varier le seuil d'attention t. Les mots avec une attention supérieure ou égale à t sont conservés comme explication. Le premier point est à* $10^{-}4$.

telles que la précision et le score f1 ne sont pas pertinentes, comme indiqué dans le tableau 2. Une IoU élevée entre les Ancres et les explications d'Attention indique que les explications sont similaires avec les deux méthodes. Dans l'ensemble, en comparaison avec la vérité terrain, les explications de l'Attention sont légèrement meilleures que les Ancres, cf. Table 2.

TAB. 2 – *Evaluation des Ancres et de l'Attention avec la vérité terrain, pour les prédictions correctes de LEGO seulement, les meilleurs résultats sont en gras.*

Mesure	**Ancres**	**Attention**	**Ancres vs Attention**
IoU	0.9377	**0.9573**	0.9471
Taux de reconnaissance	0.9803	**0.9871**	*non pertinent*
Rappel	**0.9696**	0.9641	*non pertinent*
Précision	0.9614	**0.9932**	*non pertinent*
F1-score	0.9540	**0.9688**	*non pertinent*

Pour le cas d'usage YELP, aucune vérité terrain n'est disponible. La comparaison n'est possible qu'entre les explications des Ancres et de l'Attention. L'IoU indique si les explications données sont similaires. L'IoU moyenne sur l'ensemble de test est de 0,2292, ce qui montre de fortes différences entre les deux méthodes d'explication. Cela peut s'expliquer par des textes longs et un vocabulaire vaste attendu dans les explications. Par conséquent, pour évaluer les méthodes d'explication dans le cas d'usage YELP, une analyse qualitative est nécessaire.

4.2 Analyse qualitative

Comme une IoU élevée montre une forte similitude entre deux explications, l'analyse qualitative peut être plus efficace en filtrant les textes à faible IoU dans le jeu de test. Cela indique si une méthode d'explication est plus précise lorsque des différences subsistent. Par conséquent, ce filtrage sera utilisé dans l'analyse qualitative suivante pour les deux cas d'usage.

Pour le cas d'usage LEGO, les explications des Ancres sont plus courtes que celles basées sur l'Attention. Les longueurs moyennes sont respectivement de 0,15 et 0,33 mots dans le jeu

de test. La valeur moyenne est faible à cause des explications vides. Le tableau suivant (cf. Table 3) donne quelques exemples où l'IoU est inférieure à 0,5. Cette analyse qualitative est en accord avec l'analyse quantitative et souligne que les explications à base d'Attention sont meilleures pour ce cas d'usage. Sur l'ensemble de données YELP, il n'y a aucune possibilité de comparer les explications avec la vérité terrain. Pourtant, il est intéressant d'analyser les explications des critiques extrêmes (1 et 5 étoiles, correctement classées) lorsque l'IoU vaut 0, ce qui signifie que les explications des deux méthodes comparées sont différentes.

TAB. 3 – *Textes de LEGO avec différentes explications. Le texte est au dessus des autres informations.*

Motif de rejet	**Vérité terrain**	**Ancre**	**Attention**
"Contrat a duree indeterminee - Dfd Notre agence de Saint-Medard-en-Jalles recherche une Assistante Administrative pour completer son equipe."			
Genre	['assistante administrative']	['recherche', 'Assistante', 'Jalles']	['assistante', 'administrative']
"Nous recherchons actuellement un Teleconseiller FRANCAIS / NEERLANDAIS (H/F) pour le compte de notre client, a Marcq-en-Baroeul."			
Nationalité	['francais / neerlandais']	['un', 'neerlandais', 'recherchons', 'francais']	['neerlandais']

TAB. 4 – *Textes de YELP avec differentes explications. Le texte est au dessus des autres informations.*

Etoiles	**Ancres**	**Attention**
Wow ! Superb Maids did an amazing job cleaning my house. They stayed as long as it took to make sure everything was immaculate. I will be using them on a regular basis.		
5	[]	['superb', 'amazing', 'everything']
For the record, this place is not gay friendly. Very homophobic and sad for 2019. Avoid at all costs		
1	['not']	['record', 'not', 'homophobic', 'sad', 'avoid']
Had the best experience buying my dress at brilliant bridal in jan 2018. Can't wait to wear my beautiful gown in oct 2018		
5	['brilliant']	['best', 'buying', 'can']

Les longueurs d'explication moyennes sont similaires, 2,34 et 2,13 mots respectivement pour les Ancres et l'Attention. Les deux premières lignes du tableau 4 indiquent un manque de mots significatifs pour les Ancres. Les explications basées sur l'Attention sont par conséquent le meilleur choix. La dernière ligne du tableau 4 montre des explications basées sur des mots différents mais significatifs. Les longueurs moyennes sont similaires et l'Attention semble plus précise lorsque les explications sont très différentes. Cette analyse qualitative indique que les explications basées sur l'Attention sont un choix plus sûr dans ce cas d'usage particulier.

Un point intéressant est que les deux explications soulignent les mêmes parties du texte lorsque les critiques mentionnent leur propre note, comme indiqué dans le tableau 5. Pour le premier exemple, cela conduit même à une mauvaise prédiction.

TAB. 5 – *Mentions des notes dans les explications.*

Texte	Etoiles	Prédiction	Ancre	Attention
[...] An this is the reason I gave them a mere 2 stars[...]	3	2	['2', 'stars']	['2']
3 Stars is about right. [...]	3	3	['3']	['3', 'decent']
[...] Perfect amount of sweet. 5/5 bobas.	5	5	['5/5', 'sweet', 'Perfect', 'great', 'all']	['5']

5 Conclusion

La multiplicité émergente des méthodes d'explicabilité en intelligence artificielle conduit à la nécessité de choisir une méthode qui convient à chaque cas d'usage. Dans cet article, deux cas d'usages ont été développés. L'un est disponible pour être partagé avec la communauté scientifique, et le second est extrait d'un réel besoin pour Pôle Emploi. L'utilisation de métriques comme moyen d'évaluer les explications s'est avérée très utile lorsque le jeu de test d'explications avec vérité terrain est disponible. Cependant, la création de cet ensemble de données peut être coûteuse et nécessite la contribution d'experts. Les études utilisateurs sont plus pertinentes mais sont également assez coûteuses. Pour réduire ce coût et conserver les analyses par les utilisateurs, les comparaisons de méthodes avec de fortes dissimilarités peuvent être effectuées, en s'appuyant sur l'IoU.

Un autre critère qui n'a pas été identifié à première vue est le coût de calcul et le temps de génération des explications. La génération d'explications des Ancres étant coûteuse, cela a conduit pour un cas d'usage au filtrage de l'ensemble de test. Comme l'explicabilité des algorithmes rencontre souvent l'éthique, l'IA responsable et même l'informatique écologique, on peut se demander si des méthodes d'explication basées sur l'Attention peuvent être préférées en fonction de critères d'efficience.

Enfin, comparer diverses méthodes d'explication et observer lorsque les explications sont similaires ou en désaccord peut aider à en savoir plus sur un modèle d'IA. Ce processus peut être utilisé pour évaluer le modèle lui-même.

Références

Bau, D., B. Zhou, A. Khosla, A. Oliva, et A. Torralba (2017). Network dissection : Quantifying interpretability of deep visual representations. In *Proc of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 6541–6549.

Codella, N. C., M. Hind, K. N. Ramamurthy, M. Campbell, A. Dhurandhar, K. R. Varshney, D. Wei, et A. Mojsilovic (2019). Ted : Teaching ai to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*.

Costa, F., S. Ouyang, P. Dolog, et A. Lawlor (2018). Automatic generation of natural language explanations. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, pp. 1–2.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, et L. Kagal (2018). Explaining explanations : An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, et D. Pedreschi (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR) 51*(5), 1–42.

Iyer, R., Y. Li, H. Li, M. Lewis, R. Sundar, et K. Sycara (2018). Transparency and explanation in deep reinforcement learning neural networks. In *Proc. of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Karpathy, A., J. Johnson, et F. Li (2015). Visualizing and understanding recurrent networks. *CoRR abs/1506.02078*.

Lin, Z., M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, et Y. Bengio (2017). A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Olah, C. et S. Carter (2016). Attention and augmented recurrent neural networks. *Distill*.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). "why should i trust you ?" : Explaining the predictions of any classifier. In *Proc. of the 22Nd ACM Int. Conf. on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Ribeiro, M. T., S. Singh, et C. Guestrin (2018). Anchors : High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp. 1527–1535.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, et D. Batra (2017). Grad-cam : Visual explanations from deep networks via gradient-based localization. In *The IEEE Int. Conf. on Computer Vision*.

Zeiler, M. D. et R. Fergus (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833.

Zou, Z., Z. Shi, Y. Guo, et J. Ye (2019). Object detection in 20 years : A survey. *arXiv preprint arXiv :1905.05055*.

Summary

Recent advances in eXplainable Artificial Intelligence (XAI) led to many different methods in order to improve explainability of deep learning algorithms. With many options at hand, and maybe the need to adapt existing ones to new problems, one may find in a struggle to choose the right method to generate explanations. This paper presents an objective approach to compare two different existing XAI methods. These methods are applied to a use case from literature and to a real use case of a French administration, Pôle Emploi.