Panorama des outils de visualisation pour l'explicabilité en apprentissage profond pour le traitement automatique de la langue

Alexis Delaforge, Jérôme Azé, Sandra Bringay, Arnaud Sallaberry, Maximilien Servajean

Panorama des outils de visualisation pour l'explicabilité en apprentissage profond pour le traitement automatique de la langue

A. Delaforge[1,3], J. Azé[1], S. Bringay[1,2], A. Sallaberry[1,2], M. Servajean[1,2]

[1] LIRMM, UMR 5506, Université de Montpellier, CNRS, Montpellier, France
[2] AMIS, Université Paul-Valéry, Montpellier, France
[3] Zortify Labs, Zortify, Luxembourg, Luxembourg

prenom.nom@lirmm.fr

Résumé

L'avènement de l'Intelligence Artificielle (IA) et plus spécifiquement des modèles d'apprentissage profond, a été accompagné de résultats impressionnants dans le domaine du Traitement Automatique de la Langue (TAL). Derrière les performances des réseaux de neurones se cachent de nombreuses problématiques, comme l'interprétabilité. Dans cet article d'état de l'art, nous dressons un panorama des outils de visualisation spécifiques à l'explicabilité des méthodes d'apprentissage profond en TAL qui visent à pallier le caractère boîte noire de ces approches.

Mots-clés

Interprétabilité, transparence, explicabilité, XAI, visualisation, réseaux de neurones.

Abstract

The advent of Artificial Intelligence (AI), and more specifically of deep learning models, has been accompanied by impressive results in the field of Natural Language Processing (ALP). Behind the performance of neural networks lie many issues, such as interpretability. In this state of the art article, we present an overview of visualization tools specific to the explicability of deep learning methods in NLP, which aim to overcome the black box character of these approaches.

Keywords

Interpretability, transparency, explainability, XAI, visualization, neural networks.

1 Introduction

L'avènement de l'Intelligence Artificielle (*IA*), accompagné de résultats impressionnants dans le domaine du Traitement Automatique de la Langue (*TAL*) [88] soulève des problématiques, comme l'interprétabilité [25, 97, 34], l'éthique [62, 26] et la sécurité [29, 34]. Dans cet article, nous nous intéressons aux outils pour interpréter les méthodes d'apprentissage profond en palliant leur caractère boîte noire. Autour de cette problématique, les travaux d'Hohman et al. [32] proposent un état de l'art traitant de l'utilisation de la visualisation de données dans le contexte de l'apprentissage profond. La Rosa et al. [43] proposent un état de l'art s'intéressant spécifiquement aux méthodes d'explication des prédictions. Chatzimparmpas et al. [11] proposent, eux, un état de l'art des états de l'art sur ces sujets. La principale différence avec ces trois travaux est que l'on s'intéresse dans cet article spécifiquement au *TAL* et à l'explication des prédictions en *TAL*. Gilpin et al. [23] étudient les méthodes d'explication en apprentissage profond. Similairement, Madsen et al. [55] s'intéressent à la même problématique mais pour le *TAL*. La principale différence de nos travaux avec ces derniers est que nous nous concentrons ici sur les outils de visualisation servant l'explication des prédictions pour les données textuelles.

L'interprétabilité des méthodes d'apprentissage sert à accroître la confiance des utilisateurs. Pour augmenter l'interprétabilité, il est possible : (1) d'augmenter la transparence du modèle, c'est-à-dire rendre les processus inhérents au modèle facilement compréhensibles ou (2) d'expliquer les prédictions, c'est-à-dire aider à identifier les raisons pour lesquelles un modèle a pris une décision [49].

Les méthodes d'apprentissage profond en *TAL*, nécessitent des méthodes spécifiques pour tendre vers une plus grande interprétabilité. En effet, les explications concernant des images ou des données numériques brutes se construisent différemment de celles représentant un concept contenu dans un mot capturé par un modèle d'apprentissage profond. Il est donc essentiel d'adapter les méthodes d'interprétabilité aux données textuelles et aux tâches à accomplir car les techniques de *TAL* cherchent à représenter les mots, tokens, phrases ou textes dans des espaces de représentation non compréhensibles.

Les techniques de visualisation de données offrent de nombreuses opportunités [32, 43] que nous détaillerons dans cet état de l'art. Son objectif est de dresser un panorama des outils de visualisation spécifiques à l'explicabilité des méthodes d'apprentissage profond et des méthodes de visualisation de données utilisées par ces outils, un sous domaine de l'explicabilité de l'*IA* (eXplainable Artifical Intelligence (*XAI*) en anglais). Dans la section 2, nous redéfinirons tout d'abord les concepts d'interprétabilité, de transparence et d'explicabilité puis dans la section 3, nous présenterons les outils de visualisation utilisées pour l'inter-

prétabilité pour différents modèles d'apprentissage profond avant de conclure dans la section 4. Nous donnerons également quelques perspectives sur les challenges associés à ce type d'approche pour pallier les limitations des méthodes actuelles.

2 Interprétabilité : Explicabilité et Transparence

Les notions d'interprétabilité, d'explicabilité et de transparence ne font pas consensus dans la littérature. Lipton [49] définit deux concepts qui, ensemble, définissent l'interprétabilité : la transparence et les explications post-hoc. Waltl et Vost [92] présentent la transparence et l'interprétabilité comme des sous-catégories de l'explicabilité. Beaudouin et al. [5] utilisent l'interprétabilité et l'explicabilité comme des synonymes. Enfin, Chatzimparmpas et al. [11] utilisent les définitions de Gilpin et al. [23] qui présentent l'explicabilité comme la possibilité pour un modèle de résumer les raisons de son comportement et l'interprétabilité comme la compréhension de ce qu'un modèle a fait. Face à ce manque de consensus, nous précisons, dans la suite de cette section, notre vision de la transparence et de l'explicabilité nécessaire à l'interprétabilité d'un réseau de neurones.

2.1 Transparence d'un modèle

Nous définissons la transparence comme la facilité avec laquelle un humain peut comprendre et reproduire le fonctionnement d'un modèle, indépendamment d'une prédiction. Une régression linéaire est un modèle transparent car son fonctionnement est facilement compréhensible et reproductible par un humain. Au contraire, un réseau de neurones profond dépend de l'activation de millions de neurones. Il est impossible, de comprendre précisément leur fonctionnement ou d'expliquer pourquoi telle coordonnée d'un vecteur de représentation est définie à une valeur et ce que cette valeur représente. D'après [49], la transparence d'un modèle peut être divisée en trois parties :

- **Transp.1** La compréhension globale du fonctionnement du modèle ;
- **Transp.2** La compréhension des différentes parties du modèle ;
- **Transp.3** La compréhension des mécanismes d'apprentissage et de leur convergence vers une solution optimale.

2.2 Explication d'une prédiction

L'explication d'une prédiction, ou explication post-hoc [49] ou encore explication locale [5, 55], est faite lorsque des indicateurs, issus ou non du fonctionnement d'un modèle, sont utilisés pour expliquer sa décision. Si une explication associée à une prédiction sert marginalement à interpréter un réseau, la multiplication des explications peut donner aux utilisateurs des intuitions sur le fonctionnement du modèle. Lipton [49] propose quatre catégories d'explications :

- **Exp.1** Les explications verbales justifiantes ;
- **Exp.2** Les explications locales donnant accès à des explications plus simples ne concernant qu'un sous-ensemble de l'espace de données ;
- **Exp.3** Les explications de complexité modérée présentant des comportements pour des exemples similaires ;
- **Exp.4** Les techniques de visualisation pour explorer l'espace de représentation des données ou afficher des indications sur les parties, dans les données d'entrée, qui contribuent à la prédiction.

Certaines méthodes d'explication appartiennent à plusieurs de ces catégories. Dans cet article, nous nous focalisons sur la quatrième catégorie **Exp.4**.

Indépendamment des catégories auxquelles appartient une explication, il est nécessaire de quantifier sa qualité. Pour cela, on peut imaginer trois types d'évaluations pour mesurer cette qualité : l'utilité pour une application, l'utilité générale et la fidélité [20]. L'utilité pour une application mesure l'utilité de l'explication dans le contexte dans lequel le modèle est utilisé. Dans le cas d'un modèle classifiant des textes servant d'aide à la décision, un utilisateur des explications produites est-il plus efficace qu'un utilisateur n'y ayant pas accès, lorsque l'on mesure la qualité des décisions prises. La seconde mesure possible concerne l'utilité générale évaluant la propension d'un utilisateur à choisir le meilleur modèle parmi un ensemble de modèles, prédire le comportement d'un modèle sur de nouvelles données ou identifier les données anormales dans un jeu de données. Dans le cas d'une classification de textes, un utilisateur pourrait comprendre à l'aide d'une explication, qu'une certaine combinaison de mots produit toujours la même classification. Il pourrait comprendre que cette combinaison est importante pour la prédiction et ainsi prédire le fonctionnement du modèle sur de nouvelles données contenant cette combinaison. Ce concept d'utilité générale est également à mettre en lien avec la plausibilité de l'explication [35, 66] qui détermine à quel point une explication est convaincante pour un utilisateur. Un utilisateur qui n'est pas convaincu par les explications produites pour un modèle pourrait par exemple se reporter sur un autre modèle pour lesquelles les explications seraient plus convaincantes. La fidélité mesure à quel point l'explication reflète réellement le fonctionnement d'un modèle [35]. Ces deux précédents concepts, la plausibilité et la fidélité sont décrits plus précisément dans les travaux de Jacovi et al. [35]. Ces concepts, qui ne s'opposent pas, ne s'adressent pas au même public. La plausibilité s'adresse aux personnes qui ne sont pas expertes en apprentissage automatique mais dans un domaine dans lequel on l'applique (en *TAL* par exemple) quand la fidélité s'adresse aux experts en apprentissage profond.

3 Visualisations appliquées des réseaux de neurones

La visualisation de données cherche à résumer, à mettre en lumière les caractéristiques des données pour assister les utilisateurs dans l'analyse, la recherche des informations précises, le requêtage ou la production de nouvelles données [63]. Dans la classification de Lipton [49], la quatrième catégorie **Exp.4** est dédiée aux techniques de vi-

sualisation explorant l'espace de représentation des données ou affichant des indications sur la partie, dans les données d'entrée, qui contribue à la prédiction. Néanmoins, les techniques de visualisation ne se cantonnent pas à cette catégorie (voir tableau 1). Dans cette section, nous allons donc présenter les différentes applications des techniques de visualisation de données dans le domaine des réseaux de neurones, et plus spécifiquement dans le domaine du *TAL*. Celles-ci couvrent de larges missions, comme l'aide à la transparence, à l'explicabilité ou à la présentation des résultats issus des réseaux de neurones.

Il existe de nombreux objets à visualiser pour expliquer le fonctionnement des réseaux de neurones. Hohman et al. [32] proposent dans leur état de l'art sur la visualisation de données dans l'apprentissage profond, cinq catégories d'objets à visualiser :

- **Obj.1** L'architecture des réseaux de neurones ;
- **Obj.2** Les paramètres des réseaux de neurones ;
- **Obj.3** Les unités de calcul ou couches de neurones ;
- **Obj.4** Les vecteurs de représentation des données dans des espaces à grandes dimensions ;
- **Obj.5** Les informations agrégées issues ou non du fonctionnement du modèle.

Certaines des méthodes présentées peuvent appartenir à plusieurs de ces catégories. La dernière catégorie couvre des travaux n'appartenant pas aux autres catégories.

Dans la suite de cette section, nous nous intéressons à la visualisation de ces différents objets pour différents types de réseaux de neurones (perceptrons multicouche, réseaux de neurones convolutionnels, réseaux de neurones récurrents ou réseaux auto-attentifs).

3.1 Perceptrons multicouches et Réseau neuronal convolutif

La plupart des outils se focalisent sur la visualisation de l'architecture des réseaux de neurones (**Obj.1**) pour les Perceptrons multicouches (Multi Layer Perceptron, *MLP*) et Réseau neuronal convolutif (Convolutional Neural Network, *CNN*). Tensorboard, présent dans la bibliothèque Tensorflow [1], permet de voir les matrices de données traverser les différentes couches et de connaître leurs dimensions. Les outils comme Netron [1] ou Netscope CNN Analyzer [2] présentent des stratégies très similaires à Tensorboard. Les travaux de Harley et al. [28] et de Smilkov et al. [81] (voir TensorFlow Playground [3]) permettent aussi de visualiser les architectures des réseaux de neurones simples dans une démarche pédagogique. DAX [2], présente l'influence des tokens, et à travers quels filtres ou neurones leur influence a été augmentée (qu'elle soit dans le sens de la prédiction ou non). DAX montre également quels sont les tokens qui activent le plus les filtres ou neurones concernés à l'aide de nuages de mots. Chawla et al. [12] proposent, eux, d'expliquer la participation des tokens dans une tâche de classification via un *CNN* à l'aide de fenêtres de convolutions de différentes tailles permettant d'avoir des informations sur l'influence des n-grams de tokens et donc des meilleurs paramètres pour les fenêtres de convolutions.

1. https://github.com/lutzroeder/Netron
2. http://dgschwend.github.io/netscope/quickstart.html
3. http://playground.tensorflow.org/

3.2 Réseaux de neurones récurrents

Les réseaux de neurones récurrents (Recurrent neural networks, *RNN*) ont été pendant longtemps les réseaux incontournables en *TAL* [41]. Les différents travaux en visualisation de données à propos des *RNN* sont parfois seulement adaptés à certaines architectures (*GRU* ou *LSTM*). Dans cette partie, nous traitons donc des outils de visualisation appliqués à tous les *RNN* et précisons les architectures concernées.

Pour la visualisation de l'architecture des *RNN* (**Obj.1**) Tensorboard [1] construit la visualisation d'une cellule de type *RNN* et les processus la composant. Du fait de leur caractéristique de récurrence, la visualisation a peu d'intérêt puisque chaque cellule d'un *RNN* est identique aux autres (dans le cas d'un réseau unidirectionnel avec une unique couche de *RNN*). Les paramètres des *RNN* (**Obj.2**) ne sont eux-aussi que peu visualisés pour la même raison.

Concernant les unités de calcul ou couches de neurones des *RNN* (**Obj.3**), les travaux de Karpathy et al. [42] ou Qian et al. [70] observent l'activation de la fonction mettant à jour le nouvel état caché à l'aide du précédant contexte dans un *LSTM*. Les travaux de Kadar et al. [39, 40] montrent qu'utiliser des tokens plutôt que des caractères permet de constater si les catégories lexicales et les fonctions grammaticales (et donc l'information sémantique que peuvent porter les mots), sont capturées par les *RNN*. Linzen et al. [48] montrent que les *LSTM* capturent des structures grammaticales à l'aide d'apprentissage supervisé. Là encore, ils se basent sur l'étude des états cachés de *LSTM*. LSTMVis [83] et RNNVis [61] analysent les états cachés des *RNN* à l'aide d'outils interactifs. L'utilisateur peut sélectionner une fenêtre de tokens dans la donnée d'entrée de manière à observer l'évolution de l'état caché dans cette fenêtre. Certaines méthodes utilisent des cartes de chaleur pour identifier les mots ayant le plus participé à une prédiction [45, 46, 3, 65, 64]. Il est important de noter que pour les *RNN*, les catégories **Obj.2** et **Obj.3** se confondent car on peut considérer l'état caché comme une représentation de la donnée d'entrée pour un token et non comme une unité de calcul.

En ce qui concerne les vecteurs de représentation des données dans des espaces à grandes dimensions (**Obj.4**), Qian et al. [70] visualisent à l'aide d'une *ACP* l'espace de représentation des mots construit par un *LSTM* et l'activation résultant de leur traitement pour certaines dimensions de portes d'un *LSTM*. Linzen et al. [48] montrent que les représentations des mots construites par un *LSTM* sont capables de connaître le caractère pluriel ou singulier d'un mot alors que cette information était omise pendant l'entraînement (sans le "s" pour l'anglais pas exemple) à l'aide d'une *ACP* de quelques mots sélectionnés. Li et al. [45] utilisent également une *ACP* dans l'espace de représentation des groupes de tokens pour montrer l'influence des comparatifs et des superlatifs sur la représentation d'un groupe de

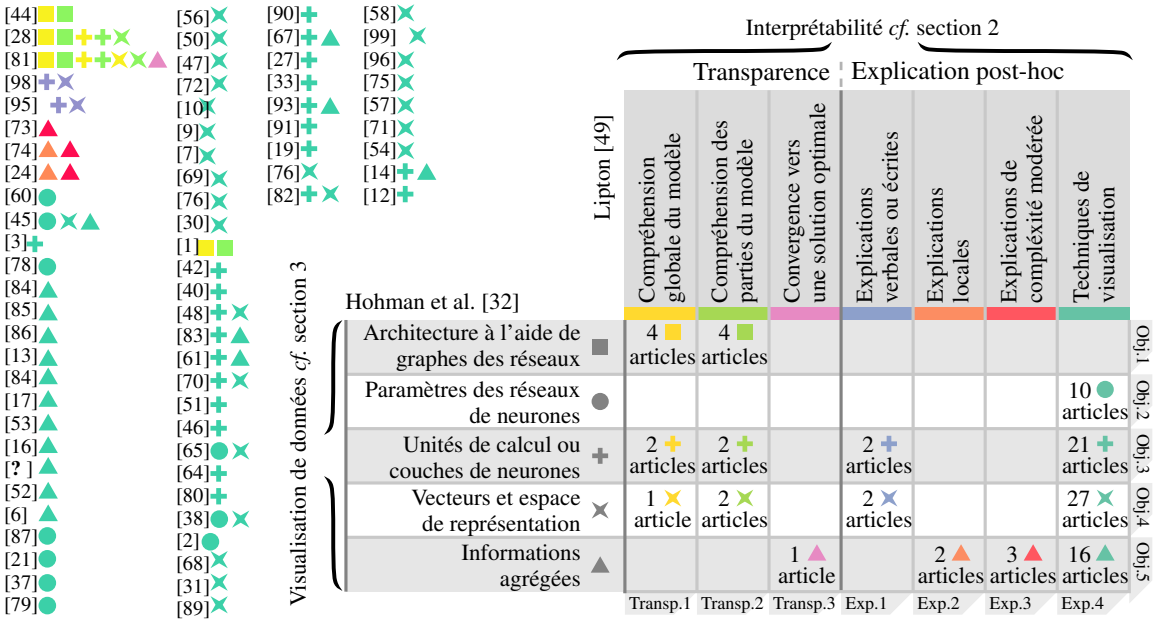

TABLE 1 – Méthodes d'interprétabilité présentées selon les classifications de Hohman et al. [32] et de Lipton [49]. Les catégories ne sont pas exhaustives car, selon le sujet, certaines catégories se confondent (*i.e.*, les unités de calcul et les vecteurs de représentation dans le cas d'utilisation de *RNN* ou réseaux auto-attentifs.)

tokens issu d'un *RNN*.

Pour visualiser des informations agrégées issues ou non du fonctionnement du modèle (**Obj.5**), les outils LSTMVis [83] et RNNVis [61] analysent des *POS* (ou étiquetage morpho-syntaxique) en plus des états cachés, par exemple en complément de la visualisation des états cachés. Li et al. [45], eux, inspectent le rôle de l'intensification et la négation dans la représentation de texte, en inspectant les états cachés des *RNN* à différents instants.

3.3 Réseaux de neurones auto-attentifs

Les réseaux de neurones auto-attentifs (dont les transformeurs) ont supplantés les *RNN* dans la plupart des tâches en *TAL*. Comme pour les *RNN*, il n'y a que peu de travaux concernant la visualisation de l'architecture (**Obj.1**) des réseaux de neurones auto-attentifs ou leurs paramètres (**Obj.2**).

Les matrices d'attention qui peuvent être vues comme des unités de calcul ou des couches de neurones (**Obj.3**), sont issues des nombreuses couches d'attention des réseaux auto-attentifs. Nlize [51] visualise des matrices d'attention pour une tâche d'inférence en langage naturel. Les cases de ces matrices sont colorées avec un encodage séquentiel des couleurs pour identifier les liens faibles ou forts d'attention entre deux tokens en fonction de la couleur foncée ou claire. Nlize visualise aussi à l'aide de graphe biparti des informations similaires. Les visualisations de graphe biparti [51, 90, 67, 14, 33] et les matrices [51, 67] sont très largement utilisés dans l'analyse de l'attention issue de transformeurs. Les visualisations de graphe biparti sont des graphiques utilisés pour comparer deux dimensions, l'une par rapport à l'autre, en utilisant une valeur de mesure comme l'encodage d'une arête entre les valeurs des dimensions. Dans le cas de l'attention, les visualisations de graphe biparti comparent un groupe de mots à un autre (le même le plus souvent) affichant comme liens entre ces deux groupes, les scores d'attention. Le plus souvent, plus ce score est grand, plus l'arête est opaque.

Vig [90] (BERTViz) ou Park et al. [67] (SANVis) proposent des visualisations de l'attention dans chacune des couches dans des transformeurs et permettent une agrégation ou une vision plus fine des têtes d'attention. Clark et al. [14] visualisent dans un espace à deux dimensions les cellules d'attention de BERT et observent les similarités de fonctionnement entre celles-ci. Hao et al. [27] s'intéressent plutôt à l'entraînement de BERT en visualisant la surface d'erreur (error surface) et démontrent que grâce au pré-entraînement et au réglage fin (fine-tuning) de BERT, celui-ci est robuste au sur-apprentissage. Ils démontrent aussi plus globalement l'efficacité d'utiliser BERT pré-entraîné. Wang et al. [93] visualisent l'attention à l'aide d'une disposition radiale des tokens. Ils visualisent aussi les différentes têtes au sein des couches pour montrer leur importance dans la tâche mais aussi la manière dont elles capturent des règles syntaxiques et sémantiques. L'analyse de l'attention ou de l'importance des cellules d'attention peut aussi se faire de manière classique. Voita et al. [91] utilisent par exemple la Layer-Wise Relevance Propagation [4] (*LRP*) pour l'analyse de l'importance des têtes d'attention dans une tâche de traduction. DeRose et al. [19] proposent AttentionFlows, un outil qui explore la façon dont l'attention des modèles auto-attentifs est affinée pendant le réglage fin, et comment l'attention informe les décisions de classification. Pour cela, ils visualisent 12 couches d'attention pour chacun des mots à l'aide

d'une visualisation radiale.
Enfin, pour visualiser les vecteurs de représentation des données (**Obj.4**) les outils s'appuient sur des travaux déjà effectués pour le plongement lexical, comme LMExplorer [76] qui présente les mots dans une réduction en deux dimensions de l'espace de représentation construit par BERT à chaque sortie des couches d'attention. Certains travaux visualisent les matrices d'attention et d'autres informations [51, 67]. Ces travaux entrent aussi dans la catégorie des outils de visualisation avec des informations agrégées (**Obj.5**).
Bien que l'attention soit beaucoup utilisée, il existe un débat autour de son usage comme méthode d'explication [8]. Jain et al. [36] avancent que l'attention n'est pas une méthode d'explication des prédictions en montrant, entre autres, qu'une modification totale des matrices d'attention de manière à ce qu'elle explique la prédiction différemment, n'a pas d'influence sur la sortie du modèle. Wiegreffe et al. [94] estiment dans leurs travaux que les expérimentations de Jain et al. [36] ne permettent pas de soutenir leur théorie. Ethayarajh et al. [22] avancent dans leur étude qu'il existe une grande similarité entre les représentations des mots dans les transformeurs (similarité cosinus). Ceci produit des matrices d'attention pour lesquelles les poids sont uniformément distribués (ceci est d'autant plus vrai pour les couches en entrée du réseau) et elle sont donc peu informatives. C'est ce que montre également les travaux de Clark et al. [14]

3.4 Méthodes agnostiques

Dans cette section, nous présentons des approches agnostiques au type de réseau. En *TAL*, il est parfois nécessaire d'utiliser des auto-encodeurs de manière à pouvoir encoder dans un espace de plus petite dimension, un long texte. Seq2Seq-Vis [82], peut être utilisé pour une tâche de traduction. En effet, l'outil affiche les scores d'attention entre un mot en entrée et un mot en sortie et affiche les espaces de représentation de ces mots mais aussi les prochains mots les plus probables pour chaque nouveau token de la traduction. L'objectif d'un tel outil est d'identifier ce qui a été appris par les auto-encodeurs et de détecter les éventuelles erreurs dans la procédure de décodage de ceux-ci pour ensuite les déboguer ou les corriger.
D'autres méthodes agnostiques permettent de visualiser les parties de la donnée d'entrée qui contribuent à la prédiction à l'aide de carte de chaleur [45, 46, 3]. Dans le cadre de l'utilisation des valeurs de Shapley [77], de SHAP [53, 52] ou de LIME [73], des diagrammes en bâtons sont également utilisés.
Certains travaux montrent l'activation précise de certaines variables des vecteurs de représentation des tokens [45]. Or, ne pouvant pas extraire l'information à propos de ce qu'est censée présenter une dimension, cela ne revêt que peu d'intérêt. Enfin, les explications hiérarchiques sont présentées sous forme d'arbre de décision montrant comment un groupe de tokens influe sur la prédiction à un instant [80, 38, 13] ou comment un modèle de substitution traite l'information à l'aide d'un arbre de décision [24].
Un dernier type de méthode permet de visualiser la frontière de décision. La visualisation de la frontière est dans le domaine de la classification de textes à l'aide de réseaux de neurones intéressante du fait de la grande dimension dans laquelle les classifieurs représentent leurs données. La visualisation de la frontière doit donc adapter ces espaces de représentation en deux ou trois dimensions. Les travaux de Migut et al. [58] ZhiyongYan et Xu [99] proposent des algorithmes pour trouver les données de la frontière de décision. Rodrigues et al. [75] visualisent la frontière de décision des classifieurs en utilisant des techniques de réduction de dimension dans l'espace d'entrée des classifieurs. Parmi les cinq techniques les plus efficaces qu'ils identifient pour visualiser la frontière de décision des *CNN* dans la classification binaire, on retrouve *UMAP* [56] et *t-SNE* [31, 89]. Cependant, l'ensemble de données utilisé dans leur expérience est linéairement séparable, ce qui n'est pas un ensemble de données réaliste et donc cette méthode n'est probablement pas efficace dans un cas réel. Les travaux précédents proposent tous des distances à la frontière qui ne sont pas fidèles à celles dans l'espace de représentation des données, ce qui ne permet pas de comparer les données entre elles en termes de confiance ou force des prédictions. Zhang et al. [96] intègrent des distances fidèles pour comparer deux classifieurs mais toute autre information sur le voisinage des données est perdue.
Dans certain cas, il est pertinent de ne s'intéresser qu'à des sous-parties d'un espace de représentation. Mikolov et al. [59], visualisent, par exemple, le lien sémantique entre pays et ville pour un sous-ensemble de mots. Abadi et al. [1] proposent, dans EmbeddingProjector[4], un échantillon de mots d'un espace de représentation. Melnik et al. [57] analysent la connectivité des données dans l'espace d'entrée. Cela garantit qu'aucune frontière de décision n'existe entre deux données si elles appartiennent aux mêmes régions de décision. Ces régions de décision sont des zones de l'espace de représentation dans lesquelles les prédictions sont toutes identiques. Différentes régions de décision sont calculées et comparées à travers différents classifieurs tels que des classifieurs neuronaux ou des *SVM* [15]. Ramamurthy et al. [71] proposent une méthodologie pour comparer, pour différents modèles, la complexité de la frontière de décision et donc la capacité de généralisation de ces modèles pour un ensemble de données. Enfin, Ma et al. [54] visualisent la décision relative à la frontière de décision en utilisant le *SVM* sur les données proches de la frontière de décision. Ils construisent plusieurs segments linéaires de la frontière de décision avec des mises en lumière de certaines parties de la frontière de décision.

4 Conclusions

Dans cet article, nous avons présenté les problématiques d'interprétabilité, de transparence et d'explicabilité et l'apport des techniques de visualisation de données combinées aux différentes architectures de réseaux de neuronnes. Le tableau 1 présente une synthèse des articles utilisant des

4. https://projector.tensorflow.org/

techniques de visualisation de données pour l'interprétabilité et leur classification dans les classifications de Lipton [49] et de Hohman [32]. Il montre que les techniques de visualisation participent entièrement au processus d'interprétabilité et pas seulement à la visualisation des espaces de représentation ou des parties de la donnée ayant participé à la prédiction.
Parmi les perspectives envisagées, nous pensons que la visualisation de la frontière de décision des réseaux de neurones est essentielle [18] car elle assiste les utilisateurs dans la compréhension des réseaux, leur débogage et la construction d'explications des prédictions. Ce problème est d'autant plus difficile dans le cadre d'une classification multi-classes. De même, la conservation des distances à la frontière dans la visualisation de l'espace de représentation, lui-même de grande dimension, est également un véritable challenge. L'exploration des localités correspondant à des parties de l'espace ajoute à cela la possibilité de comparer entre eux des exemples similaires. Pour une plus grande explicabilité des réseaux de neurones à l'aide de nouvelles visualisations, il est également envisageable de créer de nouvelles métriques d'explicabilité par exemple pour justifier des explications et générer de nouveaux exemples.

Remerciements

Ce travail a été soutenu par des subventions de la Région Occitanie [Programme "Allocation Doctorale 2019"] et du SIRIC Montpellier Cancer [Bourse INCa Inserm DGOS 12553]

Références

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow : Large-scale machine learning on heterogeneous systems, 2015.

[2] Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. Dax : Deep argumentative explanation for neural networks. *arXiv preprint :2012.05766*, 2020.

[3] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark, September 2017. ACL.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klau-schen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7) :1–46, 07 2015.

[5] Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozha-rovskyi, and Jayneel Parekh. Flexible and context-specific ai explainability : a multidisciplinary approach. *SSRN*, 2020.

[6] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. Shaff : Fast and consistent shapley effect estimates via random forests. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5563–5582. PMLR, 28–30 Mar 2022.

[7] Matthew Berger. Visually analyzing contextualized embeddings. In *2020 IEEE Visualization Conference (VIS)*, pages 276–280. IEEE, 2020.

[8] Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is attention explanation ? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 3889–3900, Dublin, Ireland, May 2022. ACL.

[9] Angie Boggust, Brandon Carter, and Arvind Satyanarayan. Embedding comparator : Visualizing differences in global structure and local neighborhoods via small multiples. In *27th International Conference on Intelligent User Interfaces*, IUI '22, page 746–766, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Gino Brunner, Yuyi Wang, Roger Wattenhofer, and Michael Weigelt. Natural language multitasking : analyzing and improving syntactic saliency of hidden representations. *arXiv preprint :1801.06024*, 2018.

[11] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3) :207–233, 2020.

[12] Piyush Chawla, Subhashis Hazarika, and Han-Wei Shen. Token-wise sentiment decomposition for convnet : Visualizing a sentiment classifier. *Visual Informatics*, 4(2) :132–141, 2020.

[13] Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint :2004.02015*, 2020.

[14] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at ? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. ACL.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 :273–297, 1995.

[16] Ian Covert and Su-In Lee. Improving kernelshap : Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR, 2021.

[17] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence : Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617, 2016.

[18] A. Delaforge, J. Aze, S. Bringay, C. Mollevi, A. Sallaberry, and M. Servajean. Ebbe-text : Explaining neural networks by exploring text classification decision boundaries. *IEEE Transactions on Visualization & Computer Graphics*, (01) :1–18, jun 5555.

[19] Joseph F DeRose, Jiayao Wang, and Matthew Berger. Attention flows : Analyzing and comparing attention mechanisms in language models. *IEEE TVCG*, 27(2) :1160–1170, 2020.

[20] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint :1702.08608*, 2017.

[21] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7) :620–631, 2021.

[22] Kawin Ethayarajh. How contextual are contextualized word representations ? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, November 2019. ACL.

[23] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations : An overview of interpretability of machine learning. In *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89, 2018.

[24] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *ArXiv*, abs/1805.10820, 2018.

[25] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5) :1–42, 2018.

[26] Thilo Hagendorff. The ethics of ai ethics : An evaluation of guidelines. *Minds and Machines*, 30(1) :99–120, 2020.

[27] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4143–4152, Hong Kong, China, November 2019. ACL.

[28] Adam W Harley. An interactive node-link visualization of convolutional neural networks. In *ISVC*, pages 867–877, 2015.

[29] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. Towards security threats of deep learning systems : A survey. *IEEE Transactions on Software Engineering*, 2020.

[30] Florian Heimerl and Michael Gleicher. Interactive analysis of word vector embeddings. In *Computer Graphics Forum*, volume 37, pages 253–265. Wiley Online Library, 2018.

[31] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, NIPS'02, pages 857–864, Cambridge, MA, USA, 2002. MIT Press.

[32] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning : An interrogative survey for the next frontiers. *IEEE TVCG*, 25(8) :2674–2693, 2019.

[33] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT : A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 187–196, Online, July 2020. ACL.

[34] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks : Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37 :100270, 2020.

[35] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems : How should we define and evaluate faithfulness ? In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Online, July 2020. ACL.

[36] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. ACL.

[37] Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations : Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104) :1–54, 2021.

[38] Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. Towards hierarchical importance attribution : Explaining compositional semantics for neural sequence models. In *International Conference on Learning Representations*, 2020.

[39] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Lingusitic analysis of multi-modal recurrent neural networks. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 8–9, Lisbon, Portugal, September 2015. ACL.

[40] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4) :761–780, 2017.

[41] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456, 2019.

[42] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint :1506.02078*, 2015.

[43] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1) :319–355, 2023.

[44] Alexander LeNail. Nn-svg : Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4(33) :747, 2019.

[45] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 681–691, San Diego, California, June 2016. ACL.

[46] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *arXiv preprint :1612.08220*, 2016.

[47] Quan Li, Kristanto Sean Njotoprawiro, Hammad Haleem, Qiaoan Chen, Chris Yi, and Xiaojuan Ma. Embeddingvis : A visual analytics approach to comparative network embedding inspection. In *2018 IEEE VAST*, pages 48–59. IEEE, 2018.

[48] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4 :521–535, 12 2016.

[49] Zachary C. Lipton. The mythos of model interpretability : In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3) :31–57, 2018.

[50] Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. Visual exploration of semantic relationships in neural word embeddings. *IEEE TVCG*, 24(1) :553–562, 2018.

[51] Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. Nlize : A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. *IEEE TVCG*, 25(1) :651–660, 2018.

[52] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv :1802.03888*, 2018.

[53] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[54] Yuxin Ma and Ross Maciejewski. Visual analysis of class separations with locally linear segments. *IEEE TVCG*, 27(1) :241–253, 2021.

[55] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp : A survey. *ACM Computing Surveys*, 55(8) :1–42, 2022.

[56] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap : Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29) :861, 2018.

[57] Ofer Melnik. Decision region connectivity analysis : A method for analyzing high-dimensional classifiers. *Machine Learning*, 48(1–3) :321–351, 2002.

[58] M. A. Migut, M. Worring, and C. J. Veenman. Visualizing multi-dimensio-nal decision boundaries in 2d. *21st ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, 29(1) :273–295, 2015.

[59] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[60] Linda Milne. Feature selection using neural networks with contribution measures. In *AI-CONFERENCE-*, pages 571–571. Citeseer, 1995.

[61] Yao Ming, Shaozu Cao, Ruixiang Zhang, Zhen Li, Yuanzhe Chen, Yangqiu Song, and Huamin Qu. Understanding hidden memories of recurrent neural networks. In *2017 IEEE VAST*, pages 13–24. IEEE, 2017.

[62] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11) :501–507, 2019.

[63] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[64] W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance : Contextual decomposition to extract interactions from lstms. *arXiv preprint :1801.05453*, 2018.

[65] W James Murdoch and Arthur Szlam. Automatic rule extraction from long short term memory networks. *arXiv preprint :1702.02540*, 2017.

[66] Duc Hau Nguyen, Guillaume Gravier, and Pascale Sébillot. A study of the plausibility of attention between rnn encoders in natural language inference. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1623–1629, 2021.

[67] Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. Sanvis : Visual analytics for understanding self-attention networks. In *2019 IEEE Visualization Conference (VIS)*, pages 146–150. IEEE, 2019.

[68] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11) :559–572, 1901.

[69] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings : Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November 2018. ACL.

[70] Peng Qian, Xipeng Qiu, and Xuan-Jing Huang. Analyzing linguistic knowledge in sequential model of sentence. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 826–835, 2016.

[71] Karthikeyan Natesan Ramamurthy, Kush Varshney, and Krishnan Mody. Topological data analysis of decision boundaries with application to model selection. In *36th International Conference on Machine Learning (ICML)*, volume 97, pages 5351–5360. PMLR, 2019.

[72] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32, 2019.

[73] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you ?" explaining the predictions of any classifier. In *22nd ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, pages 1135–1144, 2016.

[74] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors : High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[75] Francisco C. M. Rodrigues, Mateus Espadoto, Roberto Hirata, and Alexandru C. Telea. Constructing and visualizing high-quality classifier decision boundary maps. *Information*, 10(9) :280, 2019.

[76] Rita Sevastjanova, Aikaterini-Lida Kalouli, Christin Beck, Hanna Schäfer, and Mennatallah El-Assady. Explaining contextualization in language models using visual analytics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 464–476, 2021.

[77] Lloyd S Shapley. A value for n-person games, contributions to the theory of games, 2, 307–317, 1953.

[78] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[79] Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients : Feature interaction attribution for neural nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 865–878, 2021.

[80] Chandan Singh, W. James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2019.

[81] Daniel Smilkov, Shan Carter, D. Sculley, Fernanda B. Viégas, and Martin Wattenberg. Direct-manipulation visualization of deep networks. *ArXiv*, abs/1708.03788, 2017.

[82] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush. Seq2seq-vis : A visual debugging tool for sequence-to-sequence models. *IEEE TVCG*, 25(1) :353–363, 2018.

[83] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis : A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE TVCG*, 24(1) :667–676, 2017.

[84] Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11 :1–18, 2010.

[85] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3) :647–665, 2014.

[86] Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR, 2020.

[87] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[88] Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. Natural language processing advancements by deep learning : A survey. *arXiv preprint :2003.01200*, 2020.

[89] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.

[90] Jesse Vig. A multiscale visualization of attention in the transformer model. In *57th Annual Meeting of the Association for Computational Linguistics : System Demonstrations (ACL)*, pages 37–42. ACL, 2019.

[91] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention : Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019. ACL.

[92] Bernhard Waltl and Roland Vogl. Explainable artificial intelligence the new frontier in legal informatics. *Jusletter IT*, 4 :1–10, 2018.

[93] Zijie J. Wang, Robert Turko, and Duen Horng Chau. Dodrio : Exploring Transformer Models with Interactive Visualization. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing : System Demonstrations (ACL-IJCNLP)*, pages 132–141. ACL, 2021.

[94] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. ACL.

[95] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell : Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[96] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S. Ebert. Manifold : A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE TVCG*, 25(1) :364–373, 2019.

[97] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[98] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet : A semantically and visually interpretable medical image diagnosis network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6428–6436, 2017.

[99] Y. Zhiyong and Xu. Congfu. Using decision boundary to analyze classifiers. In *3rd International Conference on Intelligent System and Knowledge Engineering (ISKE)*, volume 1, pages 302–307, 2008.