Expliquer les prédictions des réseaux de neurones par l'exploration de l'espace de représentation et de la frontière de décision à l'aide d'EBBE-Text

Alexis Delaforge* Jérôme Azé* Arnaud Sallaberry*,**
Maximilien Servajean*,** Sandra Bringay*,** Caroline Mollevi***,****

*LIRMM, Université de Montpellier, CNRS
CC477 - 161 rue Ada, 4095 Montpellier Cedex 5, France
prenom.nom@lirmm.fr
http://www.lirmm.fr/
**Groupe AMIS, Université Paul-Valéry Montpellier 3
Route de Mende, 34199 Montpellier Cedex 5, France
***Institut du Cancer Montpellier (ICM)
208 Avenue des Apothicaires, Parc Euromédecine, 34298 Montpellier Cedex 5, France
caroline.mollevi@icm.unicancer.fr,
https://www.icm.unicancer.fr/fr
****Institut Desbrest d'Epidémiologie et de Santé Publique,
UMR Inserm - Université de Montpellier, Montpellier, France

Résumé. En classification automatique de textes, de nombreux travaux récents portent sur l'interprétation des réseaux de neurones par la production d'explications associées aux prédictions. Dans ce contexte, EBBE-Text offre une visualisation interactive de la frontière de décision, du positionnement des textes vis-à-vis de celle-ci (et donc de la certitude d'un réseau en ses prédictions), des chemins menant d'un texte à la frontière de décision, des informations concernant la proximité entre les textes, tout cela au sein de différentes localités dans l'espace de représentation des textes. Ces informations permettent d'intuiter comment le réseau de neurones de classification fonctionne et ainsi aider à son interprétabilité. Notre méthode crée des données sur la frontière de décision puis utilise des ensembles flous simplicials pour créer un graphe avant d'aligner linéairement les données créées sur la frontière de décision. Enfin, un processus itératif place les données d'entrée autour des arrangements linéaires des données de la frontière.

1 Introduction

Récemment, les réseaux de neurones ont connu un vif succès dans les tâches de Traitement Automatique du Langage (TAL) comme la traduction, la reconnaissance d'entités nommées ou encore l'analyse de sentiments. L'utilisation des techniques d'apprentissage profond soulèvent des questions sur l'interprétabilité et l'explicabilité de ces réseaux (Lipton, 2018). Il est, d'ailleurs, primordial de s'intéresser à ces questions, d'autant plus que le parlement européen

(Goodman et Flaxman, 2016) a fixé des règles parmi les plus strictes au monde concernant l'interprétabilité de ces réseaux de neurones.

D'après Lipton (2018), nous notons deux concepts qui constituent la définition de l'interprétabilité : la **transparence** et les **explications post-hoc**. La **transparence** se définit comme la facilité par laquelle un humain peut comprendre et reproduire le fonctionnement d'un modèle indépendamment d'une prédiction. Les **explications post-hoc** sont faites lorsque l'on se sert de différents indicateurs issus ou non du fonctionnement d'un modèle pour expliquer la prédiction qui a été faite. Des techniques en visualisation permettent de présenter l'intégralité de la structure des réseaux, les interpréter, les expliquer, les déboguer (Hohman et al., 2019). Dans ce contexte, nous proposons EBBE-Text, un outil d'aide à l'interprétabilité, offrant une visualisation de **la frontière de décision** d'un réseau de neurones. Cette visualisation de la frontière de décision et de la distance des données à celle-ci, permet une identification des prédictions et de la certitude avec laquelle le réseau les a classées. Dans nos travaux, les explications se font localement et permettent d'identifier le voisinage des données. Ce type d'approche manque actuellement aux méthodes d'explication disponibles en classification automatique de textes. En effet, ces méthodes ne proposent pas une visualisation de la distance à la frontière de décision et de la frontière de décision elle-même.

Dans cet article, nous présentons brièvement les travaux menés en interprétabilité en section 2 puis nous développons nos travaux [1] et notre méthode en section 3 avant de proposer un cas d'étude sur des données réelles en section 5.1. Enfin, nous concluons et proposons des perspectives en section 6.

2 Travaux existants

En classification de textes, différentes techniques d'explication des prédictions sont utilisables pour mettre en lumière les mots ayant le plus participé à la prédiction. Ces techniques s'appuient notamment sur le gradient (Selvaraju et al., 2020; Smilkov et al., 2017), sur les portes présentes dans les réseaux récurrents (Karpathy et al., 2015), sur la suppression de certaines dimensions des représentations apprises par les réseaux (Li et al., 2016) ou sur les mécanismes d'attention (Bahdanau et al., 2015; Vaswani et al., 2017) présents dès la construction des réseaux de neurones. Ces méthodes mettent en évidence les mots utiles à la prédiction à l'aide de cartes de chaleur. Lors de l'utilisation d'espaces de grande dimension (encodage de mots, phrases ou textes (Mikolov et al., 2013)), il est possible d'utiliser les techniques de visualisation et de réduction de dimension pour explorer ces espaces (Smilkov et al., 2016). Ces techniques (McInnes et Healy, 2018), permettant d'explorer les proximités entre les données, servent d'explications post-hoc et participent donc à l'interprétabilité.

Même s'il existe des techniques pour explorer l'espace entier des données, aucune ne permet de l'explorer à travers différentes zones, de celui-ci, contenant des proches voisins (localités) en s'intéressant à la certitude des prédictions. Les explications locales (parfois distinctes des réels mécanismes mis en jeu) ne permettent l'interprétation du réseau de neurones que si elles sont nombreuses et se complètent. Alors, un outil d'exploration globale et locale de l'espace de représentation et du comportement du réseau de neurones dans cet espace est une piste

1. http://advanse.lirmm.fr/template_container.php?template=AD/EBBE.php

prometteuse à une meilleure interprétabilité des réseaux de neurones dans les classifications automatiques et dichotomiques de textes.

3 Méthode

Une description générale de la méthode utilisée est présentée en figure 1.

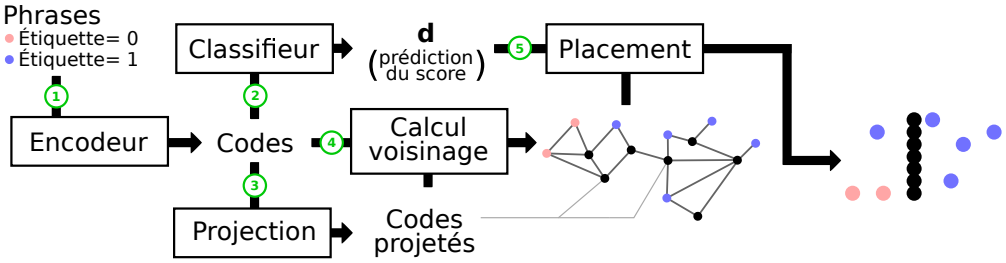

FIG. 1 – *Intégralité de la méthode de visualisation de la frontière de décision d'un réseau de neurones en classification automatique de textes. Les étapes s'appliquent selon l'ordre suivant : ①, ②, ③, ④, ⑤, ⑥. Nous décrivons les étapes d'encodage (①), de classification (②), de projection (③), de calcul du voisinage (④), de placement (⑤) en section 3.*

Pour commencer, les données sont encodées et classées à l'aide d'un réseau de neurones. Ce réseau (auto-encodeur) encode les phrases dans un espace de plus petite dimension (Hinton et Salakhutdinov, 2006) puis, à l'aide d'une régression linéaire multiple, classifie les phrases (classification dichotomique). La structure linéaire de la tâche de classification nous permet de construire des données projetées sur la frontière de décision. La frontière de décision dans l'espace de représentation est dessinée par un hyperplan. Nous construisons les vecteurs de représentation des données sur la frontière de décision en projetant orthogonalement sur l'hyperplan l'ensemble des vecteurs de représentation de nos données.

Dans l'ensemble de représentation des données et de leurs projections sur la frontière, nous calculons l'ensemble simplicial flou Zadeh (1965) associé aux données. Cette méthode est utilisée dans l'algorithme de réduction de dimension UMAP (McInnes et Healy, 2018). Pour ce faire, nous créons un ensemble flou simplicial pour chacune de ces données. Chaque donnée aura un référentiel de distance propre à elle-même fixé en fonction de la proximité avec ses voisins. Ce référentiel, étant flou, construit des probabilités de voisinage entre les données. Nous combinons ensuite tous les ensembles flous simplicials locaux en un ensemble global grâce à une union floue. Cette union des liens de voisinage entre les points permet ainsi de construire un graphe (voir figure 2). Dans ce graphe, nous scindons les composantes connexes trop grandes, avant de projeter sur une seule dimension les données projetées (sur la frontière de décision). Face à ce problème (projection sur une dimension) d'arrangement linéaire minimum, nous minimisons les distances entre les données liées à l'aide d'une stratégie proposée par Rodriguez-Tello et al. (2008).

À la suite du placement de la frontière de décision, pour chacune des composantes, nous plaçons les données (non projetées) de la plus proche à la plus éloignée de la frontière de

décision. Cette distance détermine pour chaque donnée sa position en abscisse. Pour déterminer leur position en ordonnée, nous calculons la médiane de la position de ses différents voisins déjà placés (donc plus proches de la frontière) et nous répétons cette procédure jusqu'à avoir itéré sur toutes les données d'une composante. Les données qui n'auraient pas pu être placées durant la première procédure sont placées selon la même procédure au cours des itérations suivantes. Les données se trouvant dans des composantes sans point de frontière ne sont pas affichées.

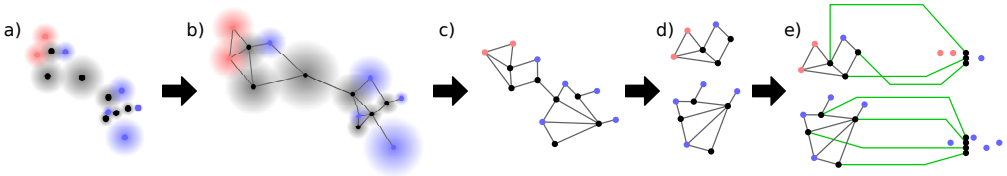

FIG. 2 – *Création du graphe et placement des données (représentées par ● ou ● selon si leur* ***étiquette*** *est 0 ou 1) et des données projetées (représentées par ●). En a) sont représentés les ensembles flous simplicials (cercles fondus) de toutes les données. Le graphe en b) est celui produit par l'union des distances entre deux données dans leurs ensembles flous respectifs. Le graphe en c) est le graphe final de toutes nos données. Le graphe en d) est celui produit à la suite de la séparation des composantes. Le graphe en e) est celui issu du placement des données projetées sur la frontière et du placement des données. Les données à gauche de la frontière de décision sont donc prédites à 0, celles à droite à 1. Les traits* ***verts*** *montrent le placement d'un point de frontière dans la visualisation finale.*

4 Outil de visualisation globale et locale des données et de la frontière de décision

L'une des caractéristiques spécifiques d'EBBE-Text est qu'il permet d'explorer l'espace de représentation des phrases à travers différentes localités (zones de l'espace de représentation). Pour parvenir à une exploration simplifiée et maximum de cette frontière de décision, nous ordonnons nos localités tout d'abord par colonne. Une visualisation de sept colonnes est produite, où, dans chaque colonne les localités sont classées de la plus grande composante à la plus petite. Enfin, les composantes sont classées au sein de ces sept colonnes en fonction de la répartition de chacune des classes dans ces composantes. Ainsi, les composantes sur la gauche de la visualisation sont celles comportant la plus grande proportion des données d'une classe alors que les composantes sur la droite sont celles comportant la plus grande proportion des données de l'autre classe. Ces techniques de classification des données ne sont possibles que lorsque les **étiquettes** sont connues (jeu d'entraînement par exemple).

Lorsqu'une localité est sélectionnée, nous proposons une visualisation de la frontière de décision (vue de gauche), des phrases de cette localité (vue centrale), de différents espaces de représentation en deux dimensions construits par des méthodes de réduction de dimension (vue de droite), des mots les plus pertinents associés à la localité (vue de droite) et des résultats de

classification du réseau de neurones pour cette localité (vue de droite). À chaque survol d'une phrase dans la visualisation de la frontière de décision, la vue de droite propose un affichage des liens de voisinages de cette donnée dans les espaces de représentation. Lors de la sélection, la vue de gauche et la vue centrale montrent les chemins menant de cette phrase jusqu'à la frontière (l'une à l'aide des points représentant les phrases, l'autre grâce aux phrases elles-mêmes). La vue centrale permet d'identifier pour chacune des phrases affichées : l'étiquette de la phrase, sa classification, l'incertitude liée à la prédiction, le nombre de phrases voisines directes plus proches de la frontière de décision. Lorsque l'on est confronté à différentes localités de l'espace de représentation des phrases, il est important de comprendre ce qui fait la particularité de chacune de ces localités. Ainsi, pour compléter la frontière de décision, un classement des mots les plus pertinents (Sievert et Shirley, 2014) est établi par localité (vue de droite), de manière à identifier ce qu'on peut y trouver. Les mots de ce classement, au survol, permettent de trouver les phrases qui, dans la visualisation de la frontière de décision et dans la liste des phrases, les contiennent.

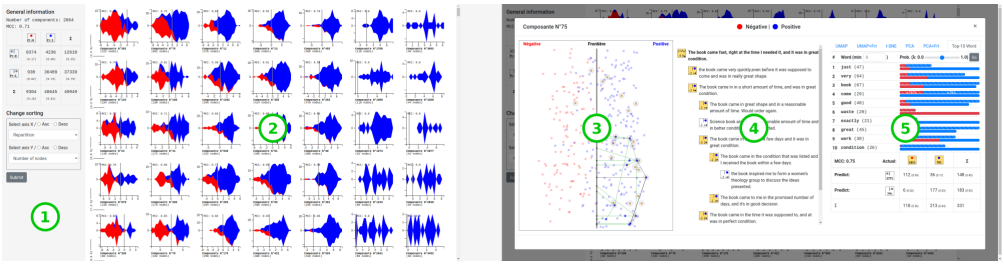

FIG. 3 – *Visualisations globale et locale. À gauche, la visualisation des différentes localités de l'espace de représentation. À droite, la visualisation d'une localité sélectionnée. En ① sont présentes des informations générales sur les classifications dans l'intégralité des données. En ②, nous proposons une abstraction des différentes localités de l'espace de représentation des phrases. En ③ la visualisation de la frontière de décision pour une localité est présentée. En ④, la phrase sélectionnée, ses voisines directes et les chemins de celle-ci à la frontière de décision sont affichés. En ⑤, se trouvent différentes visualisations d'espaces de représentation, un classement des mots les plus pertinents de la localité et des informations sur les classifications dans cette localité.*

5 Cas d'étude

5.1 Données

Le jeu de données utilisé dans notre cas d'étude est AmazonReview [2] (He et McAuley, 2016). Il comprend des avis anglophones sur différents produits vendus sur Amazon. Nous ne conservons que les avis d'au plus 20 mots étant soit les plus positifs ou les plus négatifs. Le jeu de données d'entraînement pour la classification comporte 1,3 millions d'entrées. Nos

2. http ://jmcauley.ucsd.edu/data/amazon/index_2014.html

travaux de visualisation ne portent ici que sur un échantillon de 50 000 phrases environ. Sur cet échantillon, la perplexité de la tâche d'encodage est de 1,33 et le coefficient de corrélation de Matthews de la tâche de classification est de 0,71.

FIG. 4 – *Visualisation des mots influençant la classification. En ① se trouvent les mots les plus pertinents pour cette localité. En ②, à la suite de la sélection d'une phrase on peut observer ses voisins. En ③, on peut observer la liste des phrases en lien avec celle sélectionnée. En ④, se trouvent la matrice de confusion et les informations de distribution de cette localité.*

5.2 Importance des mots

Dans la figure 4, on peut observer que la majorité des phrases contenant le mot "recevoir" (receive) sont d'**étiquette** négative et classées négativement par le réseau. Les phrases positives contenant "recevoir" sont soit sont mal classées, ou correctement classées mais avec une grande incertitude. Toujours à l'aide de cette liste de mots, nous pouvons chercher les phrases contenant le mot "encore" (yet), souvent associé en anglais au fait que l'action se soit déjà déroulée ou non. Les informations présentes dans la matrice de confusion, associées aux probabilités d'apparition de ces mots dans le jeu de données total (barre supérieure) ou dans la localité actuelle (barre inférieure), peuvent aiguiller sur le fait que les mots "recevoir" et "encore" amènent les phrases à être classées plus négativement. De plus, à l'aide d'un clic sur le mot "recevoir", on peut s'apercevoir, qu'à l'apparition de ce mot, les prédictions sont bien plus tirées vers le négatif dans cette localité que dans le reste des données. En observant que le mot

"encore" est spécifique à cette localité (au moins dix fois plus présent dans cette localité que dans le reste des données), on peut supposer qu'il y a une synergie entre les mots "recevoir" et "encore", amenant les prédictions a être classées plus négativement. Enfin, la visualisation de la frontière nous permet d'inspecter les données mal classées. Une observation du mot "recevoir" dans la phrase positive mal classée avec la plus grande certitude, nous permet de comprendre l'influence que ce mot peut avoir, et nous aiguiller sur l'influence que d'autres mots peuvent avoir. Ici, par exemple, le mot "retourner" a probablement eu de l'importance. Nous pouvons, par exemple, continuer notre exploration en s'intéressant à ce mot.

6 Conclusion

Dans cet article, nous présentons l'outil EBBE-Text. Il permet d'expliquer le fonctionnement des réseaux de neurones pour une tâche de classification automatique et dichotomique de textes à l'aide de la visualisation de la frontière de décision. Pour chaque localité de l'espace de représentation des données, EBBE-Text présente une visualisation de l'espace de représentation, un classement des mots les plus pertinents et les phrases appartenant à cette localité. L'interactivité d'EBBE-Text permet l'association des informations. Celles-ci contribuent à des traitements plus fins et à une meilleure interprétabilité des réseaux de neurones. Nos futurs travaux porteront sur l'amélioration des associations entre les informations produites par la visualisation de la frontière de décision et d'autres métriques pouvant aider à l'interprétabilité.

7 Remerciements

Ce travail a été soutenu et subventionné par la Région Occitanie [Programme "Allocation Doctorale 2019"] et le SIRIC Montpellier Cancer [Grant INCa_Inserm_DGOS_12553].

Références

Bahdanau, D., K. Cho, et Y. Bengio (2015). Neural machine translation by jointly learning to align and translate. *Computing Research Repository (CoRR) abs/1409.0473*.

Goodman, B. et S. Flaxman (2016). Eu regulations on algorithmic decision-making and a "right to explanation". *AI Magazine 38*, 50–57.

He, R. et J. McAuley (2016). Ups and downs : Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pp. 507–517.

Hinton, G. E. et R. R. Salakhutdinov (2006). Reducing the dimensionality of data with neural networks. *science 313*(5786), 504–507.

Hohman, F., M. Kahng, R. Pienta, et D. H. Chau (2019). Visual analytics in deep learning : An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics (TVCG) 25*(8), 2674–2693.

Karpathy, A., J. Johnson, et L. Fei-Fei (2015). Visualizing and understanding recurrent networks. *ArXiv abs/1506.02078*.

Li, J., W. Monroe, et D. Jurafsky (2016). Understanding neural networks through representation erasure. *ArXiv abs/1612.08220*.

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

McInnes, L. et J. Healy (2018). Umap : Uniform manifold approximation and projection for dimension reduction. *ArXiv abs/1802.03426*.

Mikolov, T., K. Chen, G. S. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. *Computing Research Repository (CoRR) abs/1301.3781*.

Rodriguez-Tello, E., J.-K. Hao, et J. Torres-Jimenez (2008). An effective two-stage simulated annealing algorithm for the minimum linear arrangement problem. *Computers & Operations Research 35*(10), 3331 – 3346. Part Special Issue : Search-based Software Engineering.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, et D. Batra (2020). Grad-cam : Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision (IJCV) 128*(2), 336–359.

Sievert, C. et K. Shirley (2014). LDAvis : A method for visualizing and interpreting topics. In *Proceedings of the ACL Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70.

Smilkov, D., N. Thorat, B. Kim, F. Viégas, et M. Wattenberg (2017). Smoothgrad : removing noise by adding noise. *ArXiv abs/1706.03825*.

Smilkov, D., N. Thorat, C. Nicholson, E. Reif, F. Viégas, et M. Wattenberg (2016). Embedding projector : Interactive visualization and interpretation of embeddings. *ArXiv abs/1611.05469*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, et I. Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 30, pp. 5998–6008. Curran Associates, Inc.

Zadeh, L. (1965). Fuzzy sets. *Information and Control 8*(3), 338 – 353.

Summary

In text classification, many recent works deal with the interpretation of neural networks by the production of prediction explanations. In this context, EBBE-Text offers an interactive visualization of the decision boundary, positionings of texts with respect to it (and thus the certainty of a network in its predictions), paths leading from a text to the decision boundary, information concerning the proximity between texts, trought different localities in the text representation space. These information make it possible to intuit how the classification neural network functions and thus helps in its interpretability. Our method creates data on the decision boundary and then uses simplistic fuzzy sets to create a graph before linearly aligning the created data on the decision boundary. Finally an iterative process places the input data around the linear arrangements of the boundary data.