Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

L'Attention est-elle de l'Explication ? Une Introduction au Débat

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang,
Thomas François, Patrick Watrin
CENTAL, IL&C, Université Catholique de Louvain
prenom.nom@uclouvain.be

RÉSUMÉ

Nous présentons un résumé en français et un résumé en anglais de l'article **Is Attention Explanation ? An Introduction to the Debate** (Bibal *et al.*, 2022), publié dans les actes de la conférence *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*.

ABSTRACT

Is Attention Explanation ? An Introduction to the Debate

We present a French abstract and an English abstract of the article (Bibal *et al.*, 2022), published in the proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022).

MOTS-CLÉS : attention, explication, apprentissage profond.

KEYWORDS: attention, explanation, deep learning.

1 Résumé en français

Les performances des modèles d'apprentissage profond dans le TAL et d'autres disciplines de l'apprentissage automatique ont conduit à une augmentation de leur popularité, et ainsi pouvoir expliquer le comportement de ces modèles devient crucial. L'attention a été envisagée comme une solution d'améliorer leur performance, tout en fournissant des explications. Cependant, un débat a émis le doute sur le pouvoir explicatif de l'attention dans les réseaux de neurones profonds. Ce débat trouve son origine dans deux articles (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Bien que le débat ait donné lieu à une littérature vaste, en raison de contributions provenant de différents domaines, le manque de dialogue entre ces travaux se fait de plus en plus ressentir.
Dans cet article, nous proposons une vue d'ensemble sur les contributions du débat en confrontant de manière critique ces différents travaux. À ces fins, nous confrontons des articles qui (i) citent les deux articles originaux et (ii) apportent des contributions significatives à la discussion. Nous articulons ce panorama de la manière suivante :

- Nous présentons les arguments initiaux apportés par les deux articles fondateurs (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019)
- Nous inspectons les revues de littérature existantes qui mentionnent le débat et montrons qu'aucune d'entre elles ne se concentre spécifiquement sur la question
- Nous structurons ensuite notre analyse de la littérature de la manière suivante :
  - Arguments en faveur de l'attention comme n'étant pas de l'explication
  - Analyses des raisons pour lesquelles l'attention n'est pas de l'explication

Volume 1 : conférence principale, pages 447–449.

    - Études sur la question de savoir si l'attention peut être de l'explication en fonction de la tâche
    - Méthodologies d'évaluation pour l'explication
    - Évaluer l'explication avec des humains
- Enfins, nous mettons en lumière les solutions, trouvées dans la littérature, visant à faire de l'attention une forme d'explication
    - Via des solutions techniques
    - et via l'inclusion d'utilisateurs dans la boucle

Cette vision holistique peut présenter un grand intérêt pour les futurs travaux de toutes les communautés concernées par ce débat. Nous résumons les défis principaux identifiés dans ces différents domaines, et concluons avec une discussion des directions les plus prometteuses sur l'attention en tant qu'explication.

2 Résumé en anglais

The performance of deep learning models in NLP and other fields of machine learning has led to a rise in their popularity, and so the need for explanations of these models becomes paramount. Attention has been seen as a solution to increase performance, while providing some explanations. However, a debate has started to cast doubt on the explanatory power of attention in neural networks. This debate originates from two seminal papers (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Although the debate has created a vast literature thanks to contributions from various areas, the lack of communication between works in those areas is becoming more and more tangible.
In this paper, we provide a clear overview of the insights on the debate by critically confronting those different works. To do so, we confront papers that (i) cite the two original papers and (ii) bring significant contributions to the discussion. We articulate the survey in the following way :

- We present the initial arguments brought by the two seminal papers (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019)
- We inspect existing surveys that mention the debate and show that none of them focus specifically on the question
- We then structure our analysis of the literature in the following way :
    - Arguments in favor of attention as not being explanation
    - Analyses of why attention is not explanation
    - Studies on whether attention can be explanation depending on the task
    - Evaluation methodologies for explanation
    - Evaluating explanations with humans
- Finally we highlight the solutions of the literature to make attention explanation
    - Via technical solutions
    - and via attention as explanation with users in the loop

This holistic vision can be of great interest for future works in all the communities concerned by this debate. We sum up the main challenges spotted in these areas, and we conclude by discussing the most promising avenues on attention as explanation.

Références

BIBAL A., CARDON R., ALFTER D., WILKENS R., WANG X., FRANÇOIS T. & WATRIN P. (2022). Is Attention Explanation ? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*.

JAIN S. & WALLACE B. C. (2019). Attention is not explanation. In *Proceedings of NAACL-HLT*, p. 3543–3556.

WIEGREFFE S. & PINTER Y. (2019). Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, p. 11–20.