Summary of "Is Attention Explanation? An Introduction to the Debate"

Zachary Canoot and Benjamin Xu
University of Texas at Dallas

Summary of paper by:
Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François∗ and Patrick Watrin∗
CENTAL, IL&C, University of Louvain, Belgium

Abstract

This document serves as a summary of the work of Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François and Patrick Watrin. Their summary focuses on the debate around attention mechanisms as a tool for explaining black box models. It is an excellent introduction to alignment research, the style of academic debate, and a good steppingstone for future research.

1 Authors and Prior Work

The lead author Adrien Babel is a BAEF Postdoctoral Fellow and graduate of the University of Colorado Anschutz Medical Campus. He has been writing papers for the last seven years. His most popular papers focus on the interpretability of machine learning models through various analysis methods. Everyone involved in the work has plenty of experience in NLP, particularly in applied linguistics. However, Thomas Francois is the most heavily cited author, and has a long history of working on translation tasks. The rest of the team either has experience in linguistics or general NLP studies that complement some of the potential use cases of this paper.

Listing each student by citation:

- Adrien Bibal has 420 citations.
- Rémi Cardon has 189 citations.
- David Alfter has 175 citations.
- Rodrigo Wilkens has 324 citations.
- Xiaoou Wang has 37 citations.
- Thomas François has 1481 citations.
- Patrick Watrin has 334 citations.

We will examine why their work is important at the end of the paper.

2 Context and the Alignment Problem

Considering that this summary is aimed at students who might know about how attention used, it would be best to provide some explanation.

The debate of attention as an explanatory tool is part of a larger push to find ways to align AI with their designers' intention. Alignment tools like the one studied in this paper solve a problem: "If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively… we had better be quite sure that the purpose put into the machine is the purpose which we really desire" (Weiner, 1960).

For example, at the core of GPT4, OpenAI's 4th iteration of their Generative Pretrained Transformer, is a self-attention mechanism. It can interpret huge bodies of text by first encoding words' positions and embedding the meaning of words or tokens in relation to each other. This encoding is basically a way to store the contextual meaning a word gains from certain words.

This stored encoding is part of a long tool chain in long transformer models of the modern day, but what if we could look at this data and discover what the model is "thinking"? We could see the implicit bias our training data has stored about certain words. The purpose of the machine, it's understanding of language or image data, would allow us to *explain* how a model makes decisions.

3 Summary

3.1 The Broad Debate

Attention layers have been considered as an explainability tool for black box models to understand their behavior. In 2019 a debate was started on if these layers represent any amount of

explainability. Note that attention layers can be calculated two different ways, either with an additive calculation or a dot-product. For the purposes of this debate, the functions are considered the same or similar.

The first to start the debate was Jain and Wallace (2019) when they found that attention does not correlate with the results of other explanatory tools. They also found that if they shuffled the attention weights, the outcome of the model could be the same. An immediate response by Wiegreffe and Pinter (2019) agreed on the first point but claimed two things. First, that shuffling the weights without retraining the model creates a fundamentally different model, and secondly that attention is only one of many possible explanations. Since then, the debate has had many additions in various contexts, summarized here for the purpose of encouraging more discussion.

Bibal et al. (2022) point out before diving into additional arguments many survey papers do not add additional arguments yet do often serve as jumping off points for discussion. Survey papers with additions to the debate only mention the debate, and the remainder of the paper serves as more of an exhaustive introduction.

3.2 Attention is Not Explanation Analysis

First, papers that add to the argument against attention as explanation are discussed. Importantly, it was found that removing important features regarding attention weights does not lead to a decision shift (Serrano and Smith, 2019). The first point of Jain and Wallace (2019) is corroborated by other papers as well for other explanation methods, like LIME and Shapley values (Thorne et al., 2019, Ethayarajh and Jurafsky, 2021)

Other papers examine why attention would not be an explanation tool. In one example that we find most telling, if you were to add random tokens to all elements in a corpus, the attention weights would allocate them to certain results even though their value for the task is negligible (Bai et al., 2021). Tutek and Šnajder (2020) also found that all input tokens have roughly the same influence on prediction when used with common RNN. They do propose a solution, that is immediately disputed by Meister et al. (2021).

3.3 Different Tasks and Evaluation Methods

In contrast to studies on attention as a local explanatory tool, some studies focused on specific tasks and how attention might be a global explanation tool. Clark et al. (2019) found that the attention heads of BERT encode syntactic information when performing syntactic dependency tagging and co-reference resolution. The same was also found by Vig and Belinkov utilizing GPT-2 (2019). There is even hope that attention is useful in non-single-sequence tasks, where attention better corresponds to feature importance (Vashishth et al., 2019).

Other studies question if correlations to other explanation tools can be used to evaluate attention as a metric. Techniques like LIME and SHAP don't always agree with each other, and manipulating weights does not always produce trained networks (Neely et al., 2021, Ju et al., 2021), which is in direct opposition to Jain and Wallace. (2019). There are even grounds that the evaluation methods used are not well defined, nor can they distinguish between if they are evaluating *faithfulness* or *plausibility* (Liu et al., 2020, Jacovi and Goldberg, 2020). In studies that focused on human-based evaluation, Jacovi and Goldberg found that attention could produce plausible explanations that aren't necessarily faithful explanations.

3.4 Discussion and Conclusion

There are two kinds of solutions that papers discuss: how we might create more faithful explanations, and how user-centered solutions make sure that attention is a plausible explanation. The solutions here are all varied and complex, involving implementing weights to counteract various problems stated above, examining hidden states, or implementing supervised attention.

There are still various approaches to discussing this topic. We can ask questions that address topics from the debate thus far:

- Is attention needed as an explanation?
- How can faithfulness and plausibility be evaluated?
- How can we measure faithfulness on its own?
- How do we establish a common ground on tasks and architecture?
- Can we further explore effective attention and weighting schemes?
- What more can be done with supervised attention?

In conclusion, Bibal et al., finds that we could combine the results of various solutions and create a supervised effective attention, as well as develop

a common ground on evaluation and concepts like faithfulness and plausibility.

4 Impact of this Paper

Before concluding, it is worth mentioning the body of work that led to this one and why it is important. We have been instructed to address the most influential body of preceding work, and in this case, Thomas François has the most citations. In his past work he worked extensively with AI as a tool to increase French readability (François and Fairon 2012) and worked on ways to practice language with computers (Bibauw, 2019).

It is intriguing to study the work of someone who has been in the field for more than twenty years and see the fruition of their labor as AI grows more powerful. It is more intriguing still to see them advise a paper like this one where its subject matter is related to simply trying to understand how powerful AI has gotten for these tasks, especially language learning. It brings to scale the importance of the problem the authors were trying to solve, and I can imagine that guidance was important to the success of this paper.

Now, did this paper really contribute to its subject matter? The paper states "The main contributions of this work are as follows:

- a summary and a discussion of the actual state of the debate by identifying convergences and disagreements in the literature;
- an extraction and structure of the main insights from papers of different areas that generally do not interact; and
- the basis for developing research on attention as explanation, with a more integrated state-of-the-art built upon a multitude of perspectives." (Bibal et al., 2022)

We believe Bibal et al. succeeded in contributing exactly what they desired to into the greater debate. Their summary was written to meet these terms, and they evaluated their work on how well it asked the question "*is attention explanation?"*. After reading this paper, it is hard to not ask the question yourself. It is even more relevant with the current state of AI, and the surrounding field of alignment research will grow ever more important as time goes on.

Acknowledgments

Thank you to Karen Mazidi for assigning this reading. It revealed some very interesting topics that had previously gone unexamined.

References

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. 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. Association for Computational Linguistics.

Norbert Weiner 1960. Some moral and technical consequences of automation. *Ideas That Created the Future,* pages 191–200.

Thomas François and Cédrick Fairon. 2012. An "AI readability" Formula for French as a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea. Association for Computational Linguistics.

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL, *Computer Assisted Language Learning*, pages 32-8 and 827-877.