Is Attention Explanation? An Introduction to the Debate

Anonymous ACL submission

Abstract

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. Although the debate has created a vast literature thanks to contributions from various areas, the lack of communication is becoming more and more tangible. In this paper, we provide a clear overview of the insights on the debate by critically confronting works from these different areas. 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 future avenues on attention as an explanation.

1 Introduction

Attention mechanisms have been widely used in various tasks of Natural Language Processing (NLP) as well as in other fields of machine learning (e.g., Computer Vision (Mnih et al., 2014; Li et al., 2019)). These mechanisms draw insight from the intuition that humans build the representation of a whole scene by dynamically focusing on relevant parts at different times (Rensink, 2000).

The general form of attention has been named differently according to authors (alignment model (Bahdanau et al., 2015) and attention mechanism (Vaswani et al., 2017)). In essence, the attention function maps a query Q and keys K to scalar scores (Vaswani et al., 2017). These scores are fed to a softmax function, in turn producing a set of attention weights that are then applied to values V. Different kinds of attention are thus possible according to how many keys are attended to (global vs. local attention, according to Luong et al. (2015)) and where the query is generated (cross vs. self-attention as in the works of Bahdanau et al. (2015) and Vaswani et al. (2017)). In this paper, we focus on attention regardless of these technical differences. There are mainly two ways of computing the attention weights $\hat{\alpha}$: Bahdanau et al. (2015) introduced additive attention $\hat{\alpha} = \text{softmax}(\mathbf{w_3}^{\text{T}}\tanh(\mathbf{W_1}K + \mathbf{W_2}Q))$, where $\mathbf{w_3}$, $\mathbf{W_1}$, $\mathbf{W_2}$ model parameters to be learned, and Vaswani et al. (2017) introduced scaled dot-product attention $\hat{\alpha} = \text{softmax}\left(\frac{KQ}{\sqrt{m}}\right)$, where m represents the dimension of K. These two forms are theoretically similar (Vaswani et al., 2017) and generally give the same results (Jain and Wallace, 2019), the dot-product form being faster on certain tasks from a practical point of view.

Since the introduction of attention mechanisms in the literature, many have seen the opportunity to use the weights for explaining neural networks (e.g., Xu et al. (2015); Martins and Astudillo (2016); Choi et al. (2016); Xie et al. (2017); Mullenbach et al. (2018)). *Explainability* in machine learning and NLP is defined as the capacity to explain a non-interpretable (Bibal and Frénay, 2016), i.e. black-box, model (Guidotti et al., 2018). The two major ways to explain black-box models are global explanations, providing clues about the behavior of the model as a whole, and local explanations, explaining particular decisions. Using attention to explain neural networks mainly pertains to the latter, even if some authors study attention for global explanation (e.g., Clark et al. (2019)).

Explanations can also be *faithful* (how close the explanation is to the inner workings of the model) (Rudin, 2019; Jacovi and Goldberg, 2020), or *plausible* (does the user consider the explanation of the model plausible?) (Riedl, 2019; Jacovi and Goldberg, 2020). It should be noted that *explanation* presupposes some degree of transparency to the user, whether it is faithful or plausible. Indeed, disregarding this aspect would entail that the most

faithful explanation is the black-box model itself.

Recently, a debate fundamentally questioned whether attention can be used as explanation (Jain and Wallace, 2019). An immediate response by Wiegreffe and Pinter (2019) challenged some of the arguments of Jain and Wallace (2019). To this day, the debate about "is attention explanation?" continues and is the source of a rich and diverse literature. Researchers from different areas have mostly contributed to this debate without referring to works outside, and sometimes even inside, their area. These insights include theoretical analyses of attention, the necessity to bring users in the loop, questioning the evaluation methodology for model explanation, and more.

This paper aims at bringing together the papers from these different areas in order to provide an outline of the quickly growing and vast literature on the subject. Moreover, we discuss the lessons learned and highlight the main issues and perspectives. To accurately reflect the debate, we only focus on papers that are posterior to the works of Jain and Wallace (2019) and Wiegreffe and Pinter (2019), and that explicitly rely on these two papers to contribute to the debate. This paper proposes the first introduction to the debate about "is attention explanation?". 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 bases for developing research on attention as explanation, with a more integrated state-of-the-art built upon a multitude of perspectives.

In order to present the different insights on the debate, we briefly summarize the two seminal papers (Section 2), describing the arguments of the two original papers that represent the source of the ongoing debate. We also present survey papers that mention the debate within a broader context (Section 3). We then investigate the different research perspectives we extracted from the literature (Sections 4 to 9). Finally, we analyze the insights offered by those works and offer foundations to build upon for future research related to attention as explanation (Section 10).

2 Starting Point of the Debate

Jain and Wallace (2019) make a set of observations on attention weights in a battery of experiments: (i) an analysis of the correlations between attention weights and feature importance methods (gradient-based and leave-one-out) and (ii) a study of the impact of counterfactual attention weight distributions on the final prediction by randomly shuffling the attention weights, and by shuffling them *adversarially* (i.e., by creating distributions that correspond to a focus on a different set of features than the one in the original attention distribution). The experiments are performed on three tasks: binary text classification, question answering and natural language inference. When commenting upon the results of their experiments, the authors' observations are: (i) there are poor correlations between attention weights and gradient-based or leave-one-out methods for explanation and (ii) shuffling the attention weights in a neural model does not affect the final prediction, except for some rare cases where the prediction relies on a few high precision tokens. The conclusion they draw from the poor correlations with other explanation methods and the lack of exclusive explanation is that attention cannot be used as a means of explanation.

Wiegreffe and Pinter (2019) agree on the importance of the questions raised by Jain and Wallace (2019) and reply to their claims. They agree with the first observation and the corresponding experimental setup. However, they object to the second claim, stating that only modifying the attention weights in the model does not produce a real attention-based model. Indeed, if the attention weights should be modified for experimental purposes, then the model should be retrained to correspond to a real trained model with those modified attention weights. In addition, they also object to the exclusive explanation argument that attention is "*an* explanation, not *the* explanation" (Wiegreffe and Pinter, 2019, p. 13). Indeed, several plausible explanations can co-exist for a similar degree of faithfulness.

The clash between the initial use of attention as explanation and the 2019 studies showing that attention might not be explanation started a vast literature on the subject. The following section presents survey papers that are mentioning the debate within a broader perspective.

3 Survey Papers Mentioning the Debate

Usually, when exploring a question, survey papers are a good starting point, as they have the advantage of covering a broader scope. However, there is no in-depth introduction to the debate, as survey papers only briefly mention the debate and sometimes do not really add something significant for the discussion (e.g., Chaudhari et al. (2019) and Lindsay (2020)). Please note that we only discuss surveys that add significant elements to the discussion.

Galassi et al. (2020) propose a survey on attention. They recall the results of Jain and Wallace (2019) on the fact that attention may not be explanation, but also refer to the fact that only *faithful* explanations (and not *plausible* ones; see Section 7) are considered. The "explanation" perspective of the survey is focused on the work of Zhang et al. (2019), which discusses how well attention captures the importance of abstract features in multi-layer neural networks when dealing with images. Galassi et al. (2020) argue that an answer to the question "is attention explanation?" with image data may not generalize to text, and should be verified, as human understanding mechanisms strongly differ between images and texts.

de Santana Correia and Colombini (2021) introduce the debate in broad terms in Section 5.7 of their survey, but point out that, based on the work of Vashishth et al. (2019), the answer to the question "is attention explanation?" can take different shapes based on the NLP task that is studied (see our Section 6 for more details on this point of the debate). Later in their paper, they also mention, like Galassi et al. (2020), that some works show that attention in transformers focuses on syntactical structures (Voita et al., 2018; Vig and Belinkov, 2019; Tenney et al., 2019; Clark et al., 2019). This indicates that global explanations based on attention can be provided, but do not answer the need for the local, decision-based, explanation that is mainly discussed in the debate.

Ras et al. (2021) also stress that the debate has been extended to several NLP tasks in the work of Vashishth et al. (2019). They add the information that mixed results have been obtained in the debate (Serrano and Smith, 2019; Baan et al., 2019).

Contrary to the short introductions to the debate in these survey papers, we aim at providing a clear and rather exhaustive view of the different ways the debate is tackled in the literature. The different insights on the debate, which are unfortunately not regrouped and discussed in these surveys (because the debate is not their main focus), are numerous: some papers add arguments about the fact that attention is not explanation (Section 4), provide analyses to explain why attention is not explanation (Section 5), analyze the debate on different NLP tasks (Section 6), discuss the methodological issues at the heart of the debate (Section 7), evaluate the explanatory power of attention with humans (Section 8), or propose solutions to make attention become explanation (based on technical developments or on user-in-the-loop strategies) (Section 9).

4 Additional Arguments about Attention is not Explanation

Some works may be considered as the direct continuation of the arguments of Jain and Wallace (2019) by adding experiments that corroborate their findings, e.g., by showing that the comparison of attention with other *explainable* methods different from the gradient-based one leads to similar conclusions.

Serrano and Smith (2019) show that removing features considered as important by attention less often leads to a decision flip than removing features considered important by gradient-based methods. This means that the features deemed important by attention for a decision are not so important for the model. This, therefore, adds to the first argument of Jain and Wallace (2019) against the relevance of attention as an indicator of feature importance.

Thorne et al. (2019) demonstrate that applying LIME (Ribeiro et al., 2016) on an attention-based neural network can provide good explanations that the attention itself cannot provide. They conclude on this subject that their experimental results are aligned with the ones of Jain and Wallace (2019).

Mohankumar et al. (2020) investigate attention on top of LSTMs (attention-LSTMs). Their study focuses on why attention in such models neither provides *plausible*, nor *faithful*, explanations. They use a variety of NLP tasks (sentiment analysis, natural language inference, question answering and paraphrase detection) and randomly permute attention weights as Jain and Wallace (2019). They find that attention-LSTM's outputs do not change much after the permutation and conclude that attention weights are not faithful explanations in attention-LSTMs. The authors propose changes to attention-LSTMs to make attention a faithful explanation (see Section 9.1). Moreover, by analyzing the attention given to part-of-speech tags, they find

that the model cannot provide a plausible explanation either, since, for several datasets, a significant amount of attention is given to punctuation.

Finally, Ethayarajh and Jurafsky (2021) show that attention weights are not Shapley values (i.e. a method for feature importance) (Lundberg and Lee, 2017). This result is in line with Jain and Wallace (2019) on the fact that the attention weights do not correlate with other explanation techniques (saliency maps or Shapley values). The authors however note that attention flows (i.e. an extension of attention weights obtained after post-processing) (Abnar and Zuidema, 2020) are Shapley values, which may indicate that using attention in another way could lead to explanation.

5 Analyses of Why Attention is not Explanation

In addition to the arguments in the literature on the fact that attention is not explanation, another part of the literature focuses on understanding the reasons *why* it is not explanation.

Bai et al. (2021) show that attention can be put on uninteresting tokens because of an effect they call "combinatorial shortcuts". The key idea is that attention is calculated on the basis of a biased input: "the attention mechanism will try to select biased features to adapt the biased estimations to minimize the overall loss functions" (Bai et al., 2021, p. 27). For instance, if one adds random tokens (such as A, B, and C) to all documents in a corpus, one might find that some of these tokens are considered as important for the positive (or negative) class because their representation ends up being similar to the representation of "good" (or "bad"), even if their information content for the task is negligible, as they are present in all documents.

Brunner et al. (2020) theoretically show that attention weights in transformers can be decomposed into two parts, from which the "effective attention" part corresponds to the attention that really affects the output. *Effective attention* focuses on the effective input needed by the model for the task and is not biased by the representation of the input. Kobayashi et al. (2020) extend the work of Brunner et al. (2020), but focus on describing the effective attention part in more detail instead of using it to improve the model. Likewise, Sun and Marasović (2021) also extend the work of Brunner et al. (2020) and delve deeper into the explanation of effective attention and its use for explaining the model.

Sun and Lu (2020) study attention through two specific scores: attention and polarization. The attention score corresponds to the absolute value associated with each input token before the transformation into an attention weight. The polarization score is a global score (not instance-specific) for each input token, indicating its importance for predicting the positive or negative class. The authors show through these two scores why attention-based models are stable in their prediction, even when attention weights differ. They also show that the match between attention and polarizing scores strongly depends on the hyperparameter values.

By analyzing the effect of regularization on attention, Tutek and Šnajder (2020) show that one of the reasons why attention cannot be used as a faithful explanation is due to the fact that all input tokens roughly have the same influence on the prediction. The authors show that regularizing attention-based models so that embedded tokens e_t better correspond to their hidden representation $rnn(e_t)$ produces explanations that are more faithful to the model. However, Meister et al. (2021) show that regularizing generally decreases the correlation between attention and explanation techniques, if the regularization is directed towards sparse attention weights. The authors conclude that sparsity, which is often viewed as increasing interpretability of models in the literature, in this case reduces the faithfulness of explanations.

Another way to analyze the problem is to study the change in the representation of the meaning of a sentence when (i) an attention layer is added, and when (ii) the type of RNN encoding the input is changed (Zhang et al., 2021). The authors show that, in addition to an increase in accuracy, the use of attention also makes the model more stable in terms of representation of sentence meanings.

6 Is Attention Explanation on Different Tasks?

In this section, we introduce arguments from the literature that claim that, despite some proofs that attention is not always explanation, attention can be explanation on certain NLP tasks. In general, attention mechanisms seem to provide faithful explanations in syntax-related tasks such as part-of-speech tagging and syntactic annotation. Clark et al. (2019) thus investigate the attention heads in BERT. They explore syntactic dependency tagging and co-reference resolution. They find that

attention heads at different layers attend to different kinds of information (e.g., direct objects of verbs, determiners of nouns or referential antecedents), with earlier layers having a broader attention span. Furthermore, attention heads in the same layer tend to show similar distributions, which is a counter to the argument of Li et al. (2018) on the fact that encouraging attention heads to learn different distributions within layers can improve performance. Overall, knowledge of syntax seems to be encoded by a variety of attention heads in different layers, and thus attention can be used as a global explanation for the tasks under investigation.

Similarly, Vig and Belinkov (2019) investigate attention in GPT-2, in particular for two tasks: part-of-speech and syntactic tagging. They find that each part-of-speech is attended to by a specific subset of attention heads, and that attention heads in adjacent layers attend to similar part-of-speech tags. In general, attention shows which tokens were attended to for the tasks at hand and can thus be used as a global explanation.

In a different vein, Vashishth et al. (2019) investigate the role of attention across a variety of NLP tasks. They show that, when the input consists of a single sequence (e.g., in sentiment classification), the attention mechanism is comparable to a gating unit and, as such, the learned weights cannot be interpreted as attention. Therefore, in this context, attention does not provide an explanation of the model's reasoning. The reduction of attention to gating units however does not hold true for self-attention networks nor for tasks depending on an additional text sequence, as for example in neural machine translation or natural language inference (pair-wise tasks and text generation tasks). In such cases, altering learned attention weights significantly degrades performance and attention appears to be an explanation of the model and to correlate with feature importance measures.

7 Evaluation Methodology for Explanation

This section focuses on critics of the methodology when evaluating explanations via attention. The critics mainly focus on two points in the evaluation setup of Jain and Wallace (2019). First, Jain and Wallace (2019) claim that there should be a consistency between attention weights and other explanation methods – which Wiegreffe and Pinter (2019) agree with – and find none. Second, they state that the fact that attention could offer different explanations (which they show by shuffling the attention weights) is an issue, which is a strong point of disagreement with Wiegreffe and Pinter (2019).

Regarding the first point, Neely et al. (2021) compare explanation methods from the literature (LIME, Integrated Gradients, DeepLIFT, Grad-SHAP and Deep-SHAP) with attention-based explanations. The comparison is performed on two types of classification: single-sequence classification (sentiment classification) and pair-sequence classification (language inference and understanding, and question answering). The authors find small agreement between the different explanation methods, including attention-based explanations. They conclude that checking for consistency between explanation methods should not be a criterion for evaluation, which goes against the agreement between the two seminal papers.

The second point on shuffling the attention weights is a subject of more discussion. Ju et al. (2021) propose a general discussion about logic traps in evaluating interpretation. Their take on this point of the debate is that a model with its manipulated attention weights in the work of Jain and Wallace (2019) "cannot even be regarded as a trained model, which makes their manipulation meaningless" (Ju et al., 2021, p. 4), which adds to the point made by Wiegreffe and Pinter (2019).

Liu et al. (2020) argue that it is too early for the debate to take place because there are no good definition and evaluation of explanations. The authors propose a Definition Driven Pipeline (DDP) to evaluate explanations based on the definition of faithfulness. They show that following this DDP can produce an evaluation of explanations that is less biased and can even drive the development of new faithful explanations.

Calling for more clearly differentiating between faithfulness and plausibility when evaluating explanation, Jacovi and Goldberg (2020) define five guidelines for evaluating faithfulness, building upon the common pitfalls and sub-optimal practices they observed in the literature. They propose an organization of the literature into three types: model assumption, prediction assumption, and linearity assumption. They state that the distinction between Jain and Wallace (2019) and Wiegreffe and Pinter (2019) is the underlying assumptions they use for evaluating attention heat-maps as explanations. The former attempts to provide differ-

ent explanations of similar decisions per instance (therefore linked to *prediction assumption*). The latter critiques the former and is more anchored in the *model assumption* type of work.

8 Evaluating Explanations with Humans

The notion of *plausibility* of attention-based explanations implies asking humans to evaluate whether attention provides a plausible explanation for the model's decisions. A first issue is whether human judges can agree on what plausible explanations of a decision (e.g., a prediction) are. In an experiment involving predictions for sentiment analysis and reading comprehension, Vashishth et al. (2019) ask humans to decide whether the top 3 highest weighted words in 200 samples are relevant for the model's prediction. They reported a very high agreement among judges (i.e. Cohen's κ over 0.8), which leads to think that words receiving the highest attention can form a plausible explanation.

A second interesting issue is the type of human annotations that should be captured in order to assess model's plausibility. The most common approach is to ask humans to assess attention heatmaps produced by a model. In Vashishth et al. (2019), users assess the relevance of the top 3 highest weighted words, whereas Mohankumar et al. (2020) ask evaluators to decide which of two attention heatmaps better explains the model's prediction as regards to three dimensions: overall prediction, completeness (which heatmap highlights all the words required for the prediction) and correctness (highlights only the important words and not unnecessary words). Another way to assess the difference between human and machine attention, in Sen et al. (2020), consists in asking humans to highlight important words for a classification task. The authors report an agreement percentage around 70% for this task and show that attention weights on top of bi-RNNs align pretty well with human attention. This finding is especially true for words for which annotators agree on the importance.

A third line of research (Sood et al., 2020) uses eye tracking measures to investigate whether machine attention match human attention. The authors hypothesize that machine attention distributions should correlate with human attention strategies for a given task (e.g., question answering). They found that human and machine attention distributions are more similar on easier tasks, which may mean that, for difficult tasks, humans required more varied strategies. For LSTMs and CNNs, diverging more from human attention leads to a drop in performance, which is not the case for XLNets.

However, the fact that humans could reliably assess model's plausibility does not ensure that the model is *faithful* (Jacovi and Goldberg, 2020). In fact, Pruthi et al. (2020) cast serious doubts on using attention maps as a way for users to audit explanations in the context of fairness. More precisely, the authors train various architectures of neural network models on datasets that are all gender-biased and whose predictions heavily rely on "impermissible" tokens (e.g., pronouns). An adapted loss function is used to penalize the attention values of these impermissible tokens. The authors conclude that, although the problematic tokens are still used by the models, they do not appear in the attention map, which wrongly leads users to believe that the models are unbiased. In other words, the authors proved that a plausible explanation does not always imply that the explanation is faithful.

9 Solutions to Make Attention Explanation

This section proposes an overview of the different solutions that have been developed to tackle the various challenges raised by the debate. We identify two types of solutions: the first type, presented in Section 9.1, concerns purely technical solutions that are often based on the theoretical and empirical analyses presented in Section 5. The second type of solutions, presented in Section 9.2, leverages user-in-the-loop strategies to align machine attention with human attention.

9.1 Technical Solutions

The technical solutions developed to make attention an explanation differ by whether they use attention values directly or indirectly. Within a recurrent network, the representation of an input element contains a summary of the components of its context. As such, the attention weight computed for that element is imprecise because it indirectly focuses on the context. In order to avoid this dispersion, some researchers seek to reinforce the link between attention weights and input elements.

Chrysostomou and Aletras (2021) produce a weighted representation of input elements using the attention weights and a score that is specific to the elements themselves. They propose three learning strategies for that score and compare their

solutions to three baseline explanations methods. Their results show that their solutions are an improvement over the baselines.

Mohankumar et al. (2020) propose the introduction of more diversity in the hidden states learned by LSTMs, allowing to observe elements separately from their context. They evaluate two different strategies and show that the resulting attention values offer explanations that are not only more faithful but also more plausible.

Tutek and Šnajder (2020) explore different hidden state regularization methods in order to preserve a strong link with the corresponding input elements. They propose a regularization scheme that positively impacts the attention weights by reinforcing their link with the model prediction, which, in turn, leads to more faithful explanations.

The above approaches rely on a property of recurrent networks and seek to work on the attention by modifying the representation of the input elements within the network. In parallel, some researchers focus directly on the attention weights with no constraints regarding the network architecture.

Moradi et al. (2021) modify the loss function by adding a term that penalizes non-faithful attention. In order to quantify faithfulness, they propose a measure that combines up to three different stress tests. They show that their method optimizes faithfulness, while improving the model's performance.

Bai et al. (2021) propose to weight the elements of the input X to counter the effect of *combinatorial shortcuts*. The weighting scheme is based on the fact that the estimation of $\mathbb{E}(Y|X \odot M)$ in attention, where M are masks applied ($\odot$) to the elements of the input X, is not the same for all elements of X.

9.2 Attention can be Explanation When Users are in the Loop

Another way to make attention become explanation is to bring users into the loop. This approach is sometimes called supervised attention, as the user attention is used by the model during training.

Strout et al. (2019) show that using human rationale to supervise attention can produce explanations that are better accepted by users, but can also lead to better results in terms of performance.

Zhong et al. (2019) modify an attention-based LSTM to make it match user provided attention. In order to do that, they compare the distributions of machine and user attention and use a Kullback–Leibler divergence between the two distributions to penalize the attention of the model.

In the same idea of supervised attention, Heo et al. (2020) extend the meta-learning technique called neural processes to include attention. Their *Neural Attention Processes* (NAP) are designed to consider user-provided attention in an active learning fashion through the use of context points.

Kanchinadam et al. (2020) also extend the training of attention to obtain a supervised version of attention. Their approach consists in the addition of a term in the objective function of their model to penalize the difference between the machine and the user attention. As in Heo et al. (2020), the authors make use of active learning in their method called *Rationale-based Active Learning with Supervised Attention* (RALSA) to collect user attention.

Finally, Arous et al. (2021) introduce *MApping human Rationales To Attention* (MARTA), a Bayesian framework to include human rationale in order to adapt machine attention. As for all other works in this section, the method improves the performance of the model while providing human-understandable explanations.

10 Discussion

As stated earlier in this paper, one of the difficulties in this debate is that the insights are brought from paper of different areas that do not always cite each other. In fact, even inside a particular area, papers do not always refer to each other. In this section, we aim at bridging the gap between the different papers and their area in order to extract the main conclusions and some points of tension.

First of all, like Thorne et al. (2019) who state that LIME can be used for explanation, thus questioning the need for attention, Bastings and Filippova (2020) state that saliency methods can be used for explanation, and so the attention is not needed in that role. Therefore, according to Bastings and Filippova (2020), if explanation tools already exist to do the job, why is the debate about attention useful? Two answers can be provided to this question. First, attention is something that is learned for performance purposes, so it would be useful if it could be used as explanation also, instead of using additional post-hoc tools. Second, the existence of the debate kick-started solutions that are now moving towards explanation.

Current solutions for making attention an explanation have to consider the two sides of explanation: faithfulness and plausibility. This subject is at the

very heart of the debate, as Wiegreffe and Pinter (2019) already mentioned the focus of Jain and Wallace (2019) on faithful explanations only. However, users may not be satisfied by explanations that are only faithful to the model, as they need to be plausible for them too. Therefore, focusing on the correlation between attention and other faithful techniques may not be enough to evaluate whether attention is explanation in real conditions. The right balance between plausibility and faithfulness may lie in human-based evaluations (Section 8) and supervised attention (Section 9.2).

That being said, faithfulness should also be evaluated on its own right, without any consideration of plausibility, to check if the explanation matches the model behavior. However, as explained by Jacovi and Goldberg (2020), faithfulness should not be evaluated in a binary fashion: the level of faithfulness needed for attention to be accepted as an explanation should be measured.

Still on the subject of evaluation, we noted that the different contributions to the debate are often based on different setups. Indeed, except for the analysis of attention on different tasks (Section 6), the contributions often base their claims on one or two tasks of their choice. We claim that a common ground must be found to properly analyze attention and its relation to explanation. The same issue has been observed with the use of different input embeddings and different architectures surrounding the attention layer(s). Likewise, Liu et al. (2020) stress that the lack of a common ground when discussing faithfulness, plausibility and explanations is not conducive to finding answers to the debate.

On the side of solutions, the common intuitive solution in interpretability and explanation that regularizing a model to be sparse improves our understanding of the model is not well supported in the literature for attention. In fact, some authors like Meister et al. (2021) note that inducing sparsity may in fact reduce the faithfulness of attention.

Another perspective that is better suited for obtaining faithful explanations is *effective attention* (Brunner et al., 2020; Kobayashi et al., 2020; Sun and Marasović, 2021). Indeed, while attention *per se* may not be explanation, further studies and uses of effective attention as a sub-part of attention may prove useful to learn a faithful explanation.

If plausible explanations, alongside faithfulness, are needed, supervised attention is a good perspective. The argument for supervised attention is well-founded: if attention is not explanation and if faithfulness is not enough, then making machine attention match human attention may be a solution. While one can argue that attention has originally been introduced for performance purposes and that supervised attention may work against this advantage, several studies show that, in fact, guiding attention increases performance (e.g., Strout et al. (2019)). Supervised attention is therefore a solution that both optimizes performance and explainability. The main cost of this solution is that it requires the participation of users, but solutions using few-shot user annotations have already been introduced in the literature (e.g., Heo et al. (2020)).

In a complementary point of view, Grimsley et al. (2020) offer a philosophical perspective on the debate. The authors show that works studying attention as explanation attempt to do so in a causal framework. They argue that it is an issue because the object of study does not fit in that type of framework. The reason is that the link between the attention layer and a model's output cannot be isolated from the other components of the model. They conclude that "attention weights alone cannot be used as causal explanation for model behavior" (Grimsley et al., 2020, p. 1786). This entails that assuming causality when evaluating the explanatory power of attention is doomed to fail by design. The authors offer other, non-causal, explanation paradigms to explore the issue, such as mathematical, structural modal, and minimal-model explanations.

11 Conclusion

We have shown that the debate about the question "is attention explanation?" already produced a vast and diverse literature. Throughout our analysis of the existing works, we have stressed various insights that could help advance the debate: theoretically refining concepts around the notion of explanation (in particular plausibility and faithfulness), developing a common ground in the evaluation setup (e.g., similar input embeddings and architectures), extending the studies and uses of effective attention, and improving the integration of users for a supervised attention. We intend that our work provides a solid ground for further research, calling for more integration to answer the question "*is attention explanation?*". In particular, combining the findings from the different areas (e.g., to produce a supervised effective attention) seems to be among the most promising avenues.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of ACL*, pages 4190–4197.

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. MARTA: Leveraging human rationales for explainable text classification. In *Proceedings of AAAI*, pages 5868–5876.

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Do transformer attention heads provide transparency in abstractive summarization? In *Proceedings of the SIGIR Workshop FACTS-IR*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why attentions may not be interpretable? In *Proceedings of the ACM SIGKDD Conference*, pages 25–34.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.

Adrien Bibal and Benoît Frénay. 2016. Interpretability of machine learning models and representations: An introduction. In *Proceedings of ESANN*, pages 77–82.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *Proceedings of ICLR*.

Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. 2019. An attentive survey of attention models. *arXiv:1904.02874*.

Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of NeurIPS*, pages 3512–3520.

George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. In *Proceedings of ACL-IJCNLP*.

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

Alana de Santana Correia and Esther Luna Colombini. 2021. Attention, please! A survey of neural attention models in deep learning. *arXiv:2103.16775*.

Kawin Ethayarajh and Dan Jurafsky. 2021. Attention flows are shapley value explanations. In *Proceedings of ACL-IJCNLP*.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2020. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4291–4308.

Christopher Grimsley, Elijah Mayfield, and Julia RS Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of LREC*, pages 1780–1790.

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

Jay Heo, Junhyeon Park, Hyewon Jeong, Kwang Joon Kim, Juho Lee, Eunho Yang, and Sung Ju Hwang. 2020. Cost-effective interactive attention learning with neural attention processes. In *Proceedings of ICML*, pages 4228–4238.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of ACL*, pages 4198–4205.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL-HLT*, pages 3543–3556.

Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. 2021. The logic traps in evaluating post-hoc interpretations. *arXiv:2109.05463*.

Teja Kanchinadam, Keith Westpfahl, Qian You, and Glenn Fung. 2020. Rationale-based human-in-the-loop via supervised attention. In *Proceedings of the KDD workshop DaSH*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of EMNLP*, pages 7057–7075.

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. In *Proceedings of EMNLP*, pages 2897–2903.

Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. 2019. Selective kernel networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 510–519.

Grace W Lindsay. 2020. Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, 14:29.

Ninghao Liu, Yunsong Meng, Xia Hu, Tie Wang, and Bo Long. 2020. Are interpretations fairly evaluated? a definition driven pipeline for post-hoc interpretability. *arXiv:2009.07494*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of NeurIPS*, pages 4768–4777.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of EMNLP*, pages 1412–1421.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of ICML*, pages 1614–1623.

Clara Meister, Stefan Lazov, Isabelle Augenstein, and Ryan Cotterell. 2021. Is sparse attention more interpretable? In *Proceedings of ACL-IJCNLP*, pages 122–129.

Volodymyr Mnih, Nicolas Heess, and Alex Graves. 2014. Recurrent models of visual attention. In *Proceedings of NeurIPS*, pages 2204–2212.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. 2020. Towards transparent and explainable attention models. In *Proceedings of ACL*, pages 4206–4216.

Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar. 2021. Measuring and improving faithfulness of attention in neural machine translation. In *Proceedings of EACL*, pages 2791–2802.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT*, pages 1101–1111.

Michael Neely, Stefan F Schouten, Maurits JR Bleeker, and Ana Lucic. 2021. Order in the court: Explainable AI methods prone to disagreement. In *Proceedings of the ICML Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of ACL*, pages 4782–4793.

Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. 2021. Explainable deep learning: A field guide for the uninitiated. *arXiv:2004.14545*.

Ronald A. Rensink. 2000. The dynamic representation of scenes. *Visual cognition*, 7(1):17–42.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD Conference*, pages 1135–1144.

Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of ACL*, pages 4596–4608.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of ACL*, pages 2931–2951.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of CoNLL*, pages 12–25.

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do human rationales improve machine explanations? In *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62.

Kaiser Sun and Ana Marasović. 2021. Effective attention sheds light on interpretability. In *Findings of ACL-IJCNLP*.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of ACL*, pages 3418–3428.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, pages 4593–4601.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. In *Proceedings of NAACL-HLT*, pages 963–969.

Martin Tutek and Jan Šnajder. 2020. Staying true to your word:(how) can attention become explanation? In *Proceedings of the ACL Workshop on Representation Learning for NLP*, pages 131–142.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, pages 1264–1274.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, pages 11–20.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of ACL*, pages 950–962.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*, pages 2048–2057.

Cheng Zhang, Qiuchi Li, Lingyu Hua, and Dawei Song. 2021. How does attention affect the model? In *Findings of ACL-IJCNLP*, pages 256–268.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *Proceedings of ICML*, pages 7354–7363.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv:1908.06870*.