Sparse attention *is* consistent with feature importance

Michael J. Neely*
Stefan F. Schouten*
michael.neely@student.uva.nl
stefan.schouten@student.uva.nl
University of Amsterdam

ABSTRACT

Attention mechanisms are commonly incorporated in models for their ability to increase performance and to provide a distribution over the inputs. The natural temptation is to assume these distributions serve as an 'explanation' of the model's decision, particularly when viewed with visualization methods such as heatmaps. In 'Attention is Not Explanation', [Jain and Wallace 2019] assert the claim made in the paper's title by empirically studying a bidirectional LSTM model with a simple attention mechanism. They offer two reasons for their claim: that attention weights do not correlate with measures of feature importance; and that attention distributions can be manipulated to produce different predictions. In this paper, we replicate their results and dispute their first claim by showing that sparse attention distributions strongly correlate with feature importance measures for the top-k features. We conclude by examining attention mechanisms in the context of explanation and recommend using other tools when making claims of interpretability.

ACM Reference Format:
Michael J. Neely and Stefan F. Schouten. 2020. Sparse attention *is* consistent with feature importance. In *Proceedings of FACT-AI*. ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Interpretability of neural networks as a research topic has gained considerable traction in recent years. [Fan et al. 2020] report an exponential increase in publications, with almost 50 publications in 2016 and over 200 in 2018. This topic is falls under the more general Transparency in AI research. Transparency concerns itself with researching how we can explain why models gave the output that they did. Attention mechanisms are frequently used to obtain such explanations, specifically through visualisation. But the validity of applying attention for this purpose, had not seen much research until recently. 'Attention is not explanation'[Jain and Wallace 2019] claims that using attention in this way is not valid.

We summarise the findings of work done after that of Jain and Wallace; study what parts of their work have come under the most scrutiny; and contribute some scrutiny of our own by showing that sparse attention could play a role in suppressing the noise that led Jain and Wallace to believe attention should not be used as explanation at all.

*Both authors contributed equally to this research.

FACT-AI, January 2020, Amsterdam, the Netherlands

2 JAIN AND WALLACE'S METHOD

'Attention is not Explanation' aims to answer two main questions: 1) "Do learned attention weights agree with alternative, natural measures of feature importance? and 2) Had we attended to different features, would the prediction have been different?"

Both questions are answered empirically. The first type of experiment (described in 2.1) aims to clarify whether or not attention agrees with alternative measures of feature importance. The second type of experiment (described in 2.2) intends to find different (counterfactual) attention weights that still obtain the same prediction. Jain and Wallace argue that these adversarial attention weights provide an equally valid explanation, and thus that the originally found set of weights are not as valuable as an explanation.

The experiments are performed on three tasks: binary text classification, question answering, and natural language inference.

2.1 Correlation with feature importance measures

Correlation is calculated with two measures of feature importance: a gradient-based method and the feature erasure (or 1leave-one-out') method. These measures are given by the following formulae:

$$g_t \leftarrow \left| \sum_{w=1}^{|V|} \mathbb{1}[\vec{x}_{tw} = 1] \frac{\partial y}{\partial \vec{x}_{tw}} \right|, \qquad \forall t \in [1, T] \tag{1}$$

$$\Delta \hat{y}_t \leftarrow \mathrm{TVD}\left(\hat{y}(\vec{x}_{-t}), \hat{y}(\vec{x})\right), \qquad \forall t \in [1, T] \tag{2}$$

Where $\hat{y}$ indicate predictions and $\vec{x}$ are the inputs, with $\vec{x}_{-t}$ as the inputs with the t-th element removed.

Kendall-τ correlation is then calculated between these measures and the attention weights.

2.2 Counterfactual attention weights

Jain and Wallace propose two methods of creating counterfactual attention weights. The first is to scramble the original attention weights. The other method is to explicitly generate so-called adversarial attention weights. These weighs are as different as possible from the original while still generating the same prediction. We will omit the formulae and algorithms here, since for reasons outlined in Section 3 and 4 our contributions do not involve this line of experimentation. In both cases, the attention weights are only altered after the model has been trained.

3 RESPONSE AND RELATED WORK

'Attention is not not Explanation' [Wiegreffe and Pinter 2019] is a direct response to Jain and Wallace which challenges multiple aspects of their paper. The main challenge is against the method of obtaining adversarial examples. They argue that "detaching the attention scores obtained by parts of the model ... degrades the model itself. The base attention weights are not assigned arbitrarily by the model, but rather computed by an integral component whose parameters were trained alongside the rest of the layers; the way they work depends on each other." To remedy this flaw, they propose an alternative way of obtaining adversarial examples. Given a base model, they train a separate model "whose explicit goal is to provide similar prediction scores for each instance, while distancing its attention distributions from [the base model]". They show that with this setup, it is still possible to find adversarial distributions, but that they are much less extreme than those found by Jain and Wallace.

Wiegreffe and Pinter also challenge the validity using some of the datasets to study attention. They compare the performance of the model to that obtained when using uniform attention weights instead of the learned ones. They find that for two of the datasets this does not influence performance. They argue this makes the observed lack of correlation irrelevant, since the weights of an unused attention mechanism have no derivable meaning.

Finally, they compare the adversarial weights with those found by training the LSTM in a setup where the weights are imposed on a simple MLP model. They also include the comparison of uniform weights with those obtained by training the MLP model. This allows for a clean comparison in a setup where, because of the model simplicity, the attention mechanism is more important. If the adversarial weights are truly equivalent explanations, they should yield similar performance in this setup as well. However, Wiegreffe and Pinter do not observe this pattern. Instead, they find that imposing adversarial weights decreases model performance considerably compared to the weights learned from training the LSTM.

'Is Attention Interpretable?' [Serrano and Smith 2019] contains an experiment similar to the one detailed in 2.1. But, instead of comparing attention mechanisms with feature importance, it compares them to the relative importance of the inputs to the attention layer itself. They find similar results, expressing their concern that "attention weights are only noisy predictors of even intermediate components' importance". We believe that this work suffers from the same problem pointed out by Wiegreffe and Pinter, because it also alters the attention weights post hoc.

'Learning to Deceive with Attention-Based Explanations' [Pruthi et al. 2019] develops a method of obtaining adversarial examples similar to that given by Wiegreffe and Pinter, using similar motivation as to why this is necessary Although they say that their results are concordant with Wiegreffe and Pinter, their conclusion has a completely different emphasis. Where Wiegreffe and Pinter mention: "We've shown that [adversarial distributions] perform poorly relative to traditional attention mechanisms when used in our diagnostic MLP model. These results indicate that trained attention mechanisms in RNNs on our datasets do in fact learn something meaningful ... which cannot be easily 'hacked' adversarially". Pruthi et al. instead say: "Amidst claims and practices that perceive attention scores to be an indication of what the model focuses on, we provide evidence that attention scores are easily manipulable".

'Attention interpretability across NLP tasks' [Vashishth et al. 2019] contains a series of experiments similar to those by Serrano and Smith. Besides setting a weight to zero and re-normalising, they also perturb the attention weights. These experiments are performed on more than just the single sequence tasks (e.g. sequence classification), they include pair sequence tasks (e.g. question answering), and generation tasks (e.g. machine translation) as well. We believe that because the perturbation happens post hoc, this too is sensitive to the same criticisms put forward by Wiegreffe and Pinter against Jain and Wallace.

[Vashishth et al. 2019] also experiment with setting the attention weights to be uniform and random. In this way, they partially reproduce the work by Wiegreffe and Pinter that labels some datasets as unusable because they do not actually use the attention mechanism. In contrast to the perturbation, these alterations do not happen entirely post hoc because they also report performance after letting the model retrain with the new attention weights. These experiments show that the tested pair sequence and generation tasks more heavily rely on their respective attention mechanisms than the single sequence tasks. Furthermore, they put forward a more fundamental argument, proposing that the attention mechanisms employed in single sequence tasks can be reduced to simple gating mechanisms. They consider the experimental results mentioned above as evidence in support of this proposition.

'AutoFocus: Interpreting Attention-based Neural Networks by Code Perturbation' [Bui et al. 2019] includes an experiment similar to the one described in 2.1. The task they perform it on is multi-class algorithm classification. In the experiment, parts of the code (statements) are removed and the effect of this on the predication is measured, similar to the leave-one-out method. This measure of importance is compared to the attention weights given by the model. They report a correlation mean of 0.65 with variance 0.26, which they consider "a strong correlation".

A different line of research into attention and transparency investigates the attention heads in multi-head attention mechanisms. Rather than looking for correlation with feature importance, they look for attention heads that seem to be performing specific NLP tasks. [Vig and Belinkov 2019] do this for the small pretrained GPT-2 model. They find that some attention heads focus on specific part-of-speech tags; and that others focus on dependency relations. Other work [Baan et al. 2019] analyses a model trained for abstractive summarisation. They observe similar things like attention heads focusing on part-of-speech tags and named-entities. But they also note that they can prune over half of the heads before seeing a significant difference in performance. The latter indicates that using heads for explanations should be done with caution, because it is unclear how important any particular head really is to the model.

4 CONTRIBUTIONS

Of the two experiments performed by Jain and Wallace, we believe that the one described in 2.2 has seen sufficient scrutiny from

Wiegreffe and Pinter. Therefore, we elect to focus our contribution on the first experiment described in 2.1.

4.1 Reproduction

We replicate the results Jain and Wallace display in the original paper using their provided code [1]. Since their experiments were already replicated by Wiegreffe and Pinter and for reasons outlined in Section 9.1, we independently reproduce their results with a new code-base. This allows us to examine Jain and Wallace's implementation in detail and demonstrate sufficient understanding of their work.

4.2 Correlation of top-k only

During the discussion of their results [Jain and Wallace 2019] note: "We ... acknowledge that irrelevant features may be contributing noise to the Kendall-τ measure, thus depressing this metric artificially". They do also argue that because the attention weights from the simpler 'averaging' model correlate much better with the other measures, noise is likely not the problem. But they go on to say: "... it remains a possibility that agreement is strong between attention weights and feature importance scores for the top-k features only (the trouble would be defining this k and then measuring correlation between non-identical sets)." This particular possibility has not been investigated by any of the work that has come out since [Jain and Wallace 2019]. In fact [Wiegreffe and Pinter 2019] mention "We find the experiments in this part of the paper convincing and do not focus our analysis here" in reference to the correlation experiment. But we do choose to investigate this. We believe attention mechanisms often assign most of the weight to just a few of the hidden states. This is also our experience with most visualisations of attention mechanisms. So if the top-k of weights *are* consistent with other measures, attention might still provide some reliable explanation.

5 EXPERIMENTAL SETUP

Like Wiegreffe and Pinter, we focus on the binary sentiment classification task with a biLSTM seq2seq encoder and linear feedforward decoder. To compare our results with the work of Jain and Wallace, we restrict our choice of datasets to the ones used for the classification task, namely: Stanford Sentiment Treebank (SST) and IMDB. Wiegreffe and Pinter show the '20News' and 'AG News' are not suitable when investigating attention, and we were unable to gain access to the MIMIC dataset. We use the same model parameters: a 128-dimensional encoder hidden state, a 300-dimensional embedding layer with pretrained FastText embeddings [Bojanowski et al. 2016], and the AMSGrad [Reddi et al. 2019] variant of the Adam [Kingma and Ba 2014] optimizer with the default PyTorch learning rate of 0.001 and ℓ_2 regularization ($\lambda = 10^{-5}$). We conduct an experiment for each combination of attention type (tanh [Bahdanau et al. 2014] or scaled dot product [Vaswani et al. 2017]), attention activation function (softmax or sparsemax) and dataset (SST or IMDB) for a total of eight experiments. We generate as outputs the same correlation graphs and correlation `.csv` files as Jain and Wallace, which display the mean, standard deviation, and significant fraction (p-value < 0.05) for each correlation by class.

[1] https://github.com/successar/AttentionExplanation/

		Gradient τ_g		Leave-One-Out τ_{loo}	
Dataset	Class	Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig Frac.
SST	0	0.280 ± 0.221	0.240	0.214 ± 0.218	0.171
	1	0.186 ± 0.262	0.240	0.174 ± 0.271	0.171
IMDB	0	0.235 ± 0.179	0.736	0.166 ± 0.144	0.646
	1	0.319 ± 0.167	0.736	0.251 ± 0.149	0.646

Table 1: reproduction mention bilstm

We conduct each experiment by training a separate model for 40 epochs. Training is terminated early if the area under the ROC curve for the validation set does not continue to improve after ten epochs. We use identical training, validation, and test splits on the datasets to those used by Jain and Wallace.

5.1 Correlation of top-k only

To solve the problem of determining an appropriate k, we propose to use sparse attention [Martins and Astudillo 2016]. We then set k equal to the amount of non-zero values in the attention weights. This avoids the problem of having to find an arbitrary value of k that works well.

The problem of measuring correlation between non-identical sets was, to our surprise, already solved by [Jain and Wallace 2019]. Their code-base contained an implementation of the top-k generalisation of Kendall-τ given in 'Comparing Top k Lists' by [Fagin et al. 2003].

6 RESULTS

6.1 Reproduction

Our results for reproducing the experiment described in 2.1 can be seen in Table 1. Our numbers show slightly lower mean correlation and slightly greater variance.

6.2 Correlation of top-k only

In Table 2 we can see the mean correlations and the corresponding standard deviations. Figure 1 shows a near identical correlation between the top-ks of the attention weights and the two feature importance measures. This image shows the correlations for the tanh attention mechanism, when k is equal to the number of non-zero entries in the. Unfortunately the results were less impressive for the sdp attention mechanism. In Figure 2 we can see that for the 'IMDB' dataset there is still some correlation, but for the 'SST' dataset there is none at all.

7 DISCUSSION

Our results confirm that when we look at a top-k only, attention can correlate just as strongly with feature importance measures as they do with each other. This is an important footnote to Jain and Wallace's main conclusion: attention might not be as precise as other metrics, but in some the circumstances we have tested, attention does seem to provide a *noisy* explanation. This is important especially in regard to visualisations of attention, as we believe most people do not judge those as if they were a complete ranking of all words in a sentence, but rather as a way to quickly look at which k words were most important to the model and to a lesser extent in what order.

Experiments in literature using attention as explanation have varied results. At the moment of writing the most that can really

				Gradient (BiLSTM) τ_g		Leave-One-Out (BiLSTM) τ_{loo}	
				tanh	sdp	tanh	sdp
Dataset	k	α	Class	Mean ± Std.	Mean ± Std.	Mean ± Std.	Mean ± Std.
SST	avg.	1 (soft)	0	0.365 ± 0.235	0.395 ± 0.261	0.316 ± 0.246	0.446 ± 0.249
		1 (soft)	1	0.312 ± 0.293	0.131 ± 0.350	0.301 ± 0.300	0.103 ± 0.366
	all	2 (sparse)	0	0.305 ± 0.228	0.258 ± 0.194	0.276 ± 0.234	0.176 ± 0.205
		2 (sparse)	1	0.336 ± 0.207	0.020 ± 0.256	0.292 ± 0.219	0.003 ± 0.255
	non-zero	2 (sparse)	0	0.824 ± 0.397	0.259 ± 0.195	0.818 ± 0.370	0.177 ± 0.207
		2 (sparse)	1	0.784 ± 0.461	0.028 ± 0.271	0.822 ± 0.356	0.012 ± 0.271
IMDB	avg.	1 (soft)	0	0.312 ± 0.169	0.166 ± 0.211	0.256 ± 0.162	0.122 ± 0.223
		1 (soft)	1	0.415 ± 0.187	0.308 ± 0.192	0.369 ± 0.171	0.320 ± 0.180
	all	2 (sparse)	0	0.248 ± 0.082	0.436 ± 0.102	0.217 ± 0.067	0.418 ± 0.113
		2 (sparse)	1	0.226 ± 0.075	0.349 ± 0.201	0.211 ± 0.072	0.305 ± 0.188
	non-zero	2 (sparse)	0	0.722 ± 0.525	0.651 ± 0.152	0.845 ± 0.346	0.621 ± 0.158
		2 (sparse)	1	0.834 ± 0.454	0.480 ± 0.176	0.910 ± 0.315	0.436 ± 0.169

Table 2: contribution

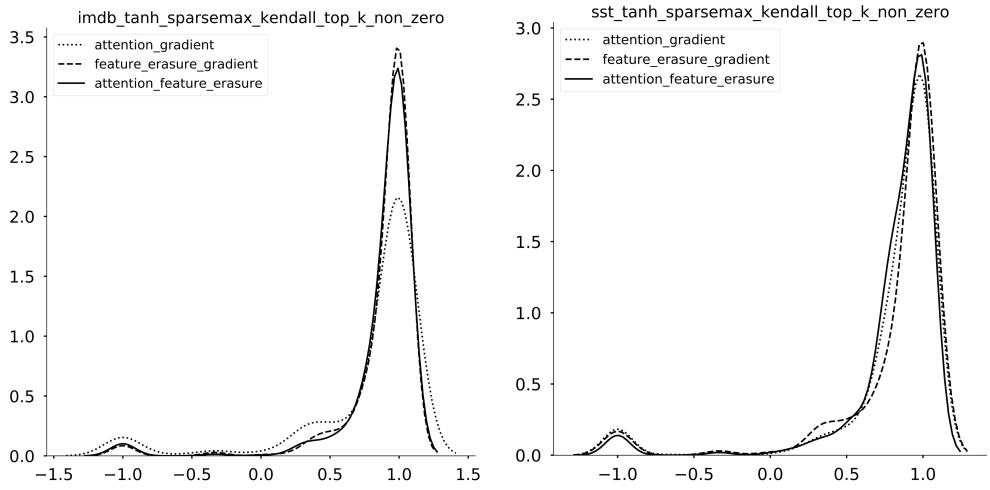

Figure 1: Density plots (for IMDB and SST) of top-k generalisation of Kendall-τ between: attention and gradient-based; attention and feature erasure; and gradient-based and feature erasure. Correlations were obtained with biLSTM model with tanh attention and with k equal to the amount of non-zero elements in the sparse attention weights.

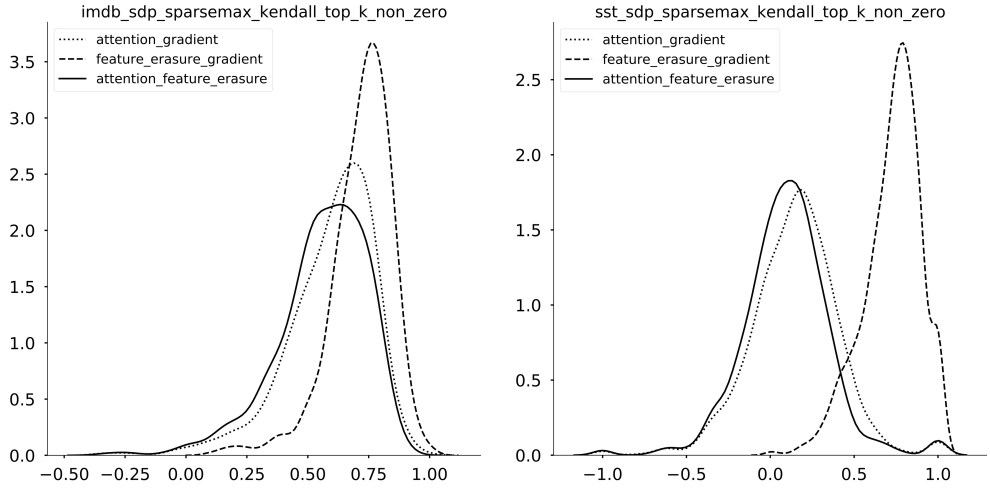

Figure 2: Same graphs as in Figure 1, but with scaled dot product attention.

be said is: attention *might* be explanation. Although no definitive conclusions have been obtained by the research carried out so far, we believe that there are some clear ways of moving forward. We know that attention does not always provide explanation. Wiegreffe and Pinter point out that the model must actually use the degrees of freedom provided by the attention mechanism, if we are to be able to actually explain anything with it. But there also seem to be variables negatively affecting the applicability of attention as explanation, without making it outright impossible. [Vashishth et al. 2019] find that the models trained for sequence-to-sequence tasks rely much more on their attention mechanisms than the models trained on classification.

From the literature it seems that the applicability of attention as explanation is influenced by: the task, the dataset and the model design. Future research should investigate this in more detail, and if there are other variables influencing this. We think there might be a correlation between: the extent to which an attention mechanism for a given dataset and model design is used; and the attention mechanism's use in explaining what the model does. In a way this is generalising what Wiegreffe and Pinter say; from a binary distinction between the attention mechanism being used or not used, to a continuous phenomenon where the attention mechanism might be used 'partially'. A way to investigate if this generalisation holds could be to plot the performance of the best adversarial examples for a dataset/model over the drop in performance measured when the attention mechanism is made defunct. If the best adversarial examples get worse as the drop in performance increases, this would indicate that the more a model depends on its attention mechanism (high performance drop) the harder it would be to find adversarial examples (worse adversarial examples).

8 BROADER IMPLICATIONS

The importance measures used in [Jain and Wallace 2019] and in this paper are relatively simple ones. A number of other more sophisticated measures are available. Measures like LIME[Ribeiro et al. 2016], Integrated Gradients[Sundararajan et al. 2017], and SHAP[Lundberg and Lee 2017]. These measures could be included in the comparison to attention, to provide a more elaborate idea of how attention compares to feature importance metrics.

9 CONCLUSION

9.1 ACM Artifact Review and Badging

Jain and Wallace provide the source code for their experiments as well as a web page for visualising their results [2]. We judge their artifacts to be *consistent*, *complete*, and *exercisable* as per the ACM[3] requirements on the 'Artifacts Evaluated' badge. However, the process of running their code to reproduce the results is non-trivial. The code is lacking in documentation and the logic is difficult to follow, let alone extend. The user is required to run multiple `iPython` notebooks and scripts to process the datasets and generate the graphs and `.csv` files. We therefore award them the 'Artifacts Evaluated – Functional' badge. To extend their experiments to the more desirable 'Artifacts Evaluated – Reusable' badge, we leverage the full features of the AllenNLP library, which was included in their code only as a scaffolding system to pass parameters to the PyTorch modules. Should anyone elect to reproduce or extend our codebase in the future they can take use all the powerful features offered by AllenNLP. Even with a limited knowledge of Python and PyTorch, users can edit the Jsonnet files to change the experiments or add new files to run different experiments. AllenNLP offers a multitude of datasets, customisable neural language model implementations, and features for visualising model behavior and results. We additionally award Jain and Wallace the 'Artifacts Available' badge since their paper, code, and results are publicly, permanently available online. To quote Wallace's reply to [Wiegreffe and Pinter 2019]: "It is nice to see research progress quickly through open science"[4].

We reproduced Jain and Wallace's Kendall tau correlation results for the binary sentiment classification task using their original code as well as our AllenNLP implementation for the IMDB and SST datasets. Additionally, Wiegreffe and Pinter successfully reproduced near-identical Classification F1 scores for all datasets. For these reasons, we award Jain and Wallace the higher-level 'Results Reproduced' badge, which requires that "the main results of the paper have been independently obtained in a subsequent study by a person or team other than the authors, without the use of author-supplied artifacts".

9.2 Attention in the context of explanation

In his examination of the 'Mythos of Model Interpretability' [Lipton 2016] highlights a common thread connecting efforts to 'explain' models: "the demand for interpretability arises when there is a mismatch between the formal objectives of supervised learning (test set predictive performance) and the real world costs in a deployment setting". Models are commonly trained with a relatively simple objective: minimize the loss between prediction and ground truth. We are often met with disappointment when we place demands of fairness and transparency on models that were not encoded in the architectural design or training process. There is a trade-off between a model's predictive power and its ability to meet human expectations of behaviour justification or intuition. Agreement exists between [Lipton 2016] and [Jain and Wallace 2019]. The former recommends "caution against blindly embracing post hoc notions of interpretability, especially when optimized to placate subjective demands", while the latter recommends "caution against using attention weights to highlight input tokens 'responsible for' model outputs and constructing just-so stories on this basis".

Attention mechanisms may not meet the strict interpretability requirements we demand of neural models. Yet, when properly framed, attention increases predictive power and seems to correlate with measures of feature importance. As such it remains to be seen if they provide honest visual justifications of their behaviour, or whether this requires something more than what even state-of-the-art feature importance metrics give us. Perhaps this is a lesson on restricting tools to their intended purpose after all. A hammer is great when you need to build a deck but it won't help you make a sandwich.

[2] https://successar.github.io/AttentionExplanation/docs/
[3] https://www.acm.org/publications/policies/artifact-review-badging
[4] https://medium.com/@byron.wallace/thoughts-on-attention-is-not-not-explanation-b7799c4c3b24

REFERENCES

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Understanding Multi-Head Attention in Abstractive Summarization. arXiv:cs.CL/1911.03898

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:cs.CL/1409.0473

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. arXiv:cs.CL/1607.04606

Nghi DQ Bui, Yijun Yu, and Lingxiao Jiang. 2019. AutoFocus: Interpreting Attention-based Neural Networks by Code Perturbation. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 38–41.

Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. 2003. Comparing top k lists. *SIAM Journal on discrete mathematics* 17, 1 (2003), 134–160.

Fenglei Fan, Jinjun Xiong, and Ge Wang. 2020. On Interpretability of Artificial Neural Networks. arXiv:cs.LG/2001.02522

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. arXiv:cs.CL/1902.10186

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv:cs.LG/1412.6980

Zachary C. Lipton. 2016. The Mythos of Model Interpretability. arXiv:cs.LG/1606.03490

Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:cs.AI/1705.07874

André F. T. Martins and Ramón Fernandez Astudillo. 2016. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. arXiv:cs.CL/1602.02068

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2019. Learning to Deceive with Attention-Based Explanations. arXiv:cs.CL/1909.07913

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the Convergence of Adam and Beyond. arXiv:cs.LG/1904.09237

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (2016). https://doi.org/10.1145/2939672.2939778

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019). https://doi.org/10.18653/v1/p19-1282

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. arXiv:cs.LG/1703.01365

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention Interpretability Across NLP Tasks. arXiv:cs.CL/1909.11218

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:cs.CL/1706.03762

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the Structure of Attention in a Transformer Language Model. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (2019). https://doi.org/10.18653/v1/w19-4808

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019). https://doi.org/10.18653/v1/d19-1002