The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?

Jasmijn Bastings
Google Research
bastings@google.com

Katja Filippova
Google Research
katjaf@google.com

Abstract

There is a recent surge of interest in using attention as explanation of model predictions, with mixed evidence on whether attention can be used as such. While attention conveniently gives us one weight per input token and is easily extracted, it is often unclear toward what goal it is used as explanation. We find that often that goal, whether explicitly stated or not, is to find out what input tokens are the most relevant to a prediction, and that the implied user for the explanation is a model developer. For this goal and user, we argue that input saliency methods are better suited, and that there are no compelling reasons to use attention, despite the coincidence that it provides a weight for each input. With this position paper, we hope to shift some of the recent focus on attention to saliency methods, and for authors to clearly state the goal and user for their explanations.

1 Introduction

Attention mechanisms (Bahdanau et al., 2015) have allowed for performance gains in many areas of NLP, including, *inter alia*, machine translation (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), natural language generation (e.g., Rush et al., 2015; Narayan et al., 2018), and natural language inference (e.g., Parikh et al., 2016).

Attention has not only allowed for better performance, it also provides a window into how a model is operating. For example, for machine translation, Bahdanau et al. (2015) visualize what source tokens the target tokens are attending to, often aligning words that are translations of each other.

Whether the window that attention gives into how a model operates amounts to *explanation* has recently become subject to debate (§2). While many papers published on the topic of explainable AI have been criticised for not defining explanations (Lipton, 2018; Miller, 2019), the first key studies which spawned interest in attention as explanation (Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019) do say that they are interested in whether attention weights faithfully represent the responsibility each input token has on a model prediction. That is, the narrow definition of explanation implied there is that it points at the most important input tokens for a prediction (arg max), accurately summarizing the reasoning process of the model (Jacovi and Goldberg, 2020b).

The above works have inspired some to find ways to make attention more faithful and/or plausible, by changing the nature of the hidden representations attention is computed over using special training objectives (e.g., Mohankumar et al., 2020; Tutek and Snajder, 2020). Others have proposed replacing the attention mechanism with a latent alignment model (Deng et al., 2018).

Interestingly, the implied definition of explanation in the cited works, happens to coincide with what *input saliency methods* (§3) are designed to produce (Li et al., 2016a; Sundararajan et al., 2017; Ribeiro et al., 2016; Montavon et al., 2019, i.a.). Moreover, the user of that explanation is often implied to be a model developer, to which faithfulness is important. The elephant in the room is therefore: If the goal of using attention as explanation is to assign importance weights to the input tokens in a faithful manner, why should the attention mechanism be preferred over the multitude of existing input saliency methods designed to do *exactly that*? In this position paper, with that goal in mind, we argue that we should pay attention no heed (§4). We propose that we reduce our focus on attention as explanation, and shift it to input saliency methods instead. However, we do emphasize that understanding the *role* of attention is still a valid research goal (§5), and finally, we discuss a few approaches that go beyond saliency (§6).

2 The Attention Debate

In this section we summarize the debate on whether attention is explanation. The debate mostly features simple BiLSTM text classifiers (see Figure 1). Unlike Transformers (Vaswani et al., 2017), they only contain a single attention mechanism, which is typically MLP-based (Bahdanau et al., 2015):

$$e_i = \mathbf{v}^\top \tanh(W_h \mathbf{h}_i + W_q \mathbf{q}) \quad \alpha_i = \frac{\exp e_i}{\sum_k \exp e_k} \quad (1)$$

where α_i is the attention score for BiLSTM state $\mathbf{h}_i$. When there is a single input text, there is no query, and $\mathbf{q}$ is either a trained parameter (like $\mathbf{v}$, W_h and W_q), or $W_q\mathbf{q}$ is simply left out of Eq. 1.

2.1 Is attention (not) explanation?

Jain and Wallace (2019) show that attention is often uncorrelated with gradient-based feature importance measures, and that one can often find a completely different set of attention weights that results in the same prediction. In addition to that, Serrano and Smith (2019) find, by modifying attention weights, that they often do not identify those representations that are most most important to the prediction of the model. However, Wiegreffe and Pinter (2019) claim that these works do not disprove the usefulness of attention as explanation *per se*, and provide four tests to determine if or when it can be used as such. In one such test, they are able to find alternative attention weights using an adversarial training setup, which suggests attention is not always a faithful explanation. Finally, Pruthi et al. (2020) propose a method to produce deceptive attention weights. Their method reduces how much weight is assigned to a set of 'impermissible' tokens, even when the models demonstratively rely on those tokens for their predictions.

2.2 Was the right task analyzed?

In the attention-as-explanation research to date text classification with LSTMs received the most scrutiny. However, Vashishth et al. (2019) question why one should focus on single-sequence tasks at all because the attention mechanism is arguably far less important there than in models involving two sequences, like NLI or MT. Indeed, the performance of an NMT model degrades substantially if uniform weights are used, while random attention weights affect the text classification performance minimally. Therefore, findings from text classification studies may not generalize to tasks where attention is a crucial component.

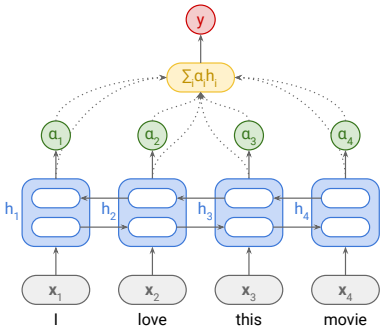

Figure 1: A typical model in the debate.

Interestingly, even for the task of MT, the first case where attention was visualized to inspect a model (§1), Ding et al. (2019) find that saliency methods (§3) yield better word alignments.

2.3 Is a causal definition assumed?

Grimsley et al. (2020) go as far as saying that attention is not explanation by definition, if a causal definition of explanation is assumed. Drawing on the work in philosophy, they point out that causal explanations presuppose that a surgical intervention is possible which is not the case with deep neural networks: one cannot intervene on attention while keeping all the other variables invariant.

2.4 Can attention be improved?

The problems with using as attention as explanation, especially regarding faithfulness, have inspired some to try and 'improve' the attention weights, so to make them more faithful and/or plausible. Mohankumar et al. (2020) observe high similarity between the hidden representations of LSTM states and propose a diversity-driven training objective that makes the hidden representations more diverse across time steps. They show using representation erasure that the resulting attention weights result in decision flips more easily as compared to vanilla attention. With a similar motivation, Tutek and Snajder (2020) use a word-level objective to achieve a stronger connection between hidden states and the words they represent, which affects attention. Not part of the recent debate, Deng et al. (2018) propose variational attention as an alternative to the soft attention of Bahdanau et al. (2015), arguing that the latter is not *alignment*, only an approximation thereof. They have the additional benefit of allowing posterior alignments, conditioned on the input and the output sentences.

3 Saliency Methods

In this section we discuss various input saliency methods for NLP as alternatives to attention: gradient-based (§3.1), propagation-based (§3.2), and occlusion-based methods (§3.3), following Arras et al. (2019). We do not endorse any specific method[1], but rather try to give an overview of methods and how they differ. We discuss methods that are applicable to *any* neural NLP model, allowing access to model internals, such as activations and gradients, as attention itself requires such access. We leave out more expensive methods that use a surrogate model, e.g., LIME (Ribeiro et al., 2016).

3.1 Gradient-based methods

While used earlier in other fields, Li et al. (2016a) use gradients as explanation in NLP and compute:

$$\nabla_{\mathbf{x}_i} f_c(\mathbf{x}_{1:n}) \tag{2}$$

where $\mathbf{x}_i$ is the input word embedding for time step i, $\mathbf{x}_{1:n} = \langle \mathbf{x}_1, \ldots, \mathbf{x}_n \rangle$ are the input embeddings (e.g., a sentence), and $f_c(\mathbf{x}_{1:n})$ the model output for target class c. After taking the L2 norm of Eq. 2, the result is a measure of how sensitive the model is to the input at time step i.

If instead we take the dot product of Eq. 2 with the input word embedding $\mathbf{x}_i$, we arrive at the gradient×input method (Denil et al., 2015), which returns a saliency (scalar) of input i:

$$\nabla_{\mathbf{x}_i} f_c(\mathbf{x}_{1:n}) \cdot \mathbf{x}_i \tag{3}$$

Integrated gradients (IG) (Sundararajan et al., 2017) is a gradient-based method which deals with the problem of *saturation*: gradients may get close to zero for a well-fitted function. IG requires a baseline $\mathbf{b}_{1:n}$, e.g., all-zeros vectors or repeated [MASK] vectors. For input i, we compute:

$$\frac{1}{m}\sum_{k=1}^{m} \nabla_{\mathbf{x}_i} f_c\Big(\mathbf{b}_{1:n} + \frac{k}{m}(\mathbf{x}_{1:n} - \mathbf{b}_{1:n})\Big) \cdot (\mathbf{x}_i - \mathbf{b}_i) \tag{4}$$

That is, we average over m gradients, with the inputs to f_c being linearly interpolated between the baseline and the original input $\mathbf{x}_{1:n}$ in m steps. We then take the dot product of that averaged gradient with the input embedding $\mathbf{x}_i$ minus the baseline.

We propose distinguishing *sensitivity* from *saliency*, following Ancona et al. (2019): the former says how much a change in the input changes the output, while the latter is the marginal effect of each input word on the prediction. Gradients measure sensitivity, whereas gradient×input and IG measure saliency. A model can be sensitive to the input at a time step, but it depends on the actual input vector if it was important for the prediction.

3.2 Propagation-based methods

Propagation-based methods (Landecker et al., 2013; Bach et al., 2015; Arras et al., 2017, i.a.), of which we discuss Layer-wise Relevance Propagation (LRP) in particular, start with a forward pass to obtain the output $f_c(\mathbf{x}_{1:n})$, which is the top-level *relevance*. They then use a special backward pass that, at each layer, *redistributes* the incoming relevance among the inputs of that layer. Each kind of layer has its own propagation rules. For example, there are different rules for feed-forward layers (Bach et al., 2015) and LSTM layers (Arras et al., 2017). Relevance is redistributed until we arrive at the input layers. While LRP requires implementing a custom backward pass, it does allow precise control to preserve relevance, and it has been shown to work better than using gradient-based methods on text classification (Arras et al., 2019).

3.3 Occlusion-based methods

Occlusion-based methods (Zeiler and Fergus, 2014; Li et al., 2016b) compute input saliency by occluding (or erasing) input features and measuring how that affects the model. Intuitively, erasing unimportant features does not affect the model, whereas the opposite is true for important features. Li et al. (2016b) erase word embedding dimensions and whole words to see how doing so affects the model. They compute the importance of a word *on a dataset* by averaging over how much, for each example, erasing that word caused a difference in the output compared to not erasing that word.

As a saliency method, however, we can apply their method on a single example only. For input i:

$$f_c(\mathbf{x}_{1:n}) - f_c(\mathbf{x}_{1:n|\mathbf{x}_i=0}) \tag{5}$$

computes saliency, where $\mathbf{x}_{1:n|\mathbf{x}_i=0}$ indicates that input word embedding $\mathbf{x}_i$ was zeroed out, while the other inputs were unmodified. Kádár et al. (2017) and Poerner et al. (2018) use a variant, *omission*, by simply leaving the word out of the input.

This method requires $n + 1$ forward passes. It is also used for evaluation, to see if important words another method has identified bring a change in model output (e.g., DeYoung et al., 2020).

[1] For an evaluation of methods for explaining LSTM-based models, see e.g., Poerner et al. (2018) and Arras et al. (2019).

4 Saliency vs. Attention

We discussed the use of attention as explanation (§2) and input saliency methods as alternatives (§3). We will now argue why saliency methods should be preferred over attention for explanation.

In many of the cited papers, whether implicitly or explicitly, the *goal* of the explanation is to reveal which input words are the most important ones for the final prediction. This is perhaps a consequence of attention computing one weight per input, so it is necessarily understood in terms of those inputs.

The intended *user* for the explanation is often not stated, but typically that user is a model developer, and not a non-expert end user, for example. For model developers, faithfulness, the need for an explanation to accurately represent the reasoning of the model, is a key concern. On the other hand, plausibility is of lesser concern, because a model developer aims to understand and possibly improve the model, and that model does not necessarily align with human intuition (see Jacovi and Goldberg, 2020b, for a detailed discussion of the differences between faithfulness and plausibility).

With this goal and user clearly stated, it is impossible to make an argument in favor of using attention as explanation. Input saliency methods are addressing the goal head-on: they reveal why one particular model prediction was made in terms of how relevant each input word was to that prediction. Moreover, input saliency methods typically take the entire computation path into account, all the way from the input word embeddings to the target output prediction value. Attention weights do not: they reflect, at one point in the computation, how much the model attends to each input *representation*, but those representations might already have mixed in information from other inputs. Ironically, attention-as-explanation is sometimes evaluated by comparing it against gradient-based measures, which again begs the question why we wouldn't use those measures in the first place.

One might argue that attention, despite its flaws, is easily extracted and computationally efficient. However, it only takes one line in a framework like TensorFlow to compute the gradient of the output w.r.t. the input word embeddings, so implementation difficulty is not a strong argument. In terms of efficiency, it is true that for attention only a forward pass is required, but many other methods discussed at most require a forward and then a backward pass, which is still extremely efficient.

5 Attention is not not interesting

In this position paper we criticized the use of attention to assess input saliency for the benefit of the model developer. We emphasize that understanding the *role* of the attention mechanism is a perfectly justified research goal. For example, Voita et al. (2019) and Michel et al. (2019) analyze the role of attention heads in the Transformer architecture and identify a few distinct functions they have, and Strubell et al. (2018) train attention heads to perform dependency parsing, adding a linguistic bias.

We also stress that if the definition of explanation is adjusted, for example if a different intended *user* and a different explanatory *goal* are articulated, attention may become a useful explanation for a certain application. For example, Strout et al. (2019) demonstrate that supervised attention helps humans accomplish a task faster than random or unsupervised attention, for a user and goal that are very different from those implied in §2.

6 Is Saliency the Ultimate Answer?

Beyond saliency. While we have argued that saliency methods are a good fit for our goal, there are other goals for which different methods can be a better fit. For example, counterfactual analysis might lead to insights, aided by visualization tools (Vig, 2019; Hoover et al., 2020; Abnar and Zuidema, 2020). The DiffMask method of DeCao et al. (2020) adds another dimension: it not only reveals in what layer a model knows what inputs are important, but also where important information is stored as it flows through the layers of the model. Other examples are models that rationalize their predictions (Lei et al., 2016; Bastings et al., 2019), which can guarantee faithful explanations, although they might be sensitive to so-called *trojans* (Jacovi and Goldberg, 2020a).

Limitations of saliency. A known problem with occlusion-based saliency methods as well as erasure-based evaluation of any input saliency technique (Bach et al., 2015; DeYoung et al., 2020) is that changes in the predicted probabilities may be due to the fact that the corrupted input falls off the manifold of the training data (Hooker et al., 2019). That is, a drop in probability can be explained by the input being OOD and not by an important feature missing. It has also been demonstrated that at least some of the saliency methods are not reliable and produce unintuitive results (Kindermans et al.,

2017) or violate certain axioms (Sundararajan et al., 2017).

A more fundamental limitation is the expressiveness of input saliency methods. Obviously, a bag of per-token saliency weights can be called an explanation only in a very narrow sense. One can overcome some limitations of the flat representation of importance by indicating dependencies between important features (for example, Janizek et al. (2020) present an extension of IG which explains pairwise feature interactions) but it is hardly possible to fully understand why a deep non-linear model produced a certain prediction by only looking at the input tokens.

7 Conclusion

We summarized the debate on whether attention is explanation, and observed that the goal for explanation is often to determine what inputs are the most relevant to the prediction. The user for that explanation often goes unstated, but is typically assumed to be a model developer. With this goal and user clearly stated, we argued that input saliency methods—of which we discussed a few—are better suited than attention. We hope, at least for the goal and user that we identified, that the focus shifts from attention to input saliency methods, and perhaps to entirely different methods, goals, and users.

Acknowledgments

We would like to thank Sebastian Gehrmann for useful comments and suggestions, as well as our anonymous reviewers, one of whom mentioned there is a whale in the room as well.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2019. *Gradient-Based Attribution Methods*, pages 169–191. Springer International Publishing, Cham.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Leila Arras, Ahmed Osman, Klaus-Robert Müller, and Wojciech Samek. 2019. Evaluating recurrent neural network explanations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 113–126, Florence, Italy. Association for Computational Linguistics.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Nicola DeCao, Michael Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9712–9724. Curran Associates, Inc.

Misha Denil, Alban Demiraj, and Nando de Freitas. 2015. Extraction of salient sentences from labelled documents.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Christopher Grimsley, Elijah Mayfield, and Julia R.S. Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of The 12th Language Resources and Evaluation Conference*,

pages 1780–1790, Marseille, France. European Language Resources Association.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9737–9748. Curran Associates, Inc.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020a. Aligning faithful interpretations with their social attribution.

Alon Jacovi and Yoav Goldberg. 2020b. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2020. Explaining explanations: Axiomatic feature interactions for deep networks. *arXiv preprint arXiv:2002.04138*.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2017. The (un)reliability of saliency methods.

W. Landecker, M. D. Thomure, L. M. A. Bettencourt, M. Mitchell, G. T. Kenyon, and S. P. Brumby. 2013. Interpreting individual classifications of hierarchical networks. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 32–38.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. Understanding neural networks through representation erasure.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14014–14024. Curran Associates, Inc.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, Online. Association for Computational Linguistics.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: An overview. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 193–209. Springer.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods*

in Natural Language Processing, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

Nina Poerner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 340–350, Melbourne, Australia. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

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

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Martin Tutek and Jan Snajder. 2020. Staying true to your word: (how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 131–142, Online. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.