Order in the Court: Explainable AI Methods Prone to Disagreement

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic

University of Amsterdam

michael@mneely.tech
sfschouten@gmail.com
m.j.r.bleeker@uva.nl
a.lucic@uva.nl

Abstract

In Natural Language Processing, feature-additive explanation methods quantify the independent contribution of each input token towards a model's decision. By computing the rank correlation between attention weights and the scores produced by a small sample of these methods, previous analyses have sought to either invalidate or support the role of attention-based explanations as a faithful and plausible measure of salience. To investigate what measures of rank correlation can reliably conclude, we comprehensively compare feature-additive methods, including attention-based explanations, across several neural architectures and tasks. In most cases, we find that none of our chosen methods agree. Therefore, we argue that rank correlation is largely uninformative and does not measure the quality of feature-additive methods. Additionally, the range of conclusions a practitioner may draw from a single explainability algorithm are limited.

1 Introduction

Of the many possible explanations for a model's decision, only those simultaneously *plausible* to human stakeholders and *faithful* to the model's reasoning process are desirable (Jacovi and Goldberg, 2020). The rest are irrelevant in the best case and harmful in the worst, particularly in critical domains such as law (Kehl and Kessler, 2017), finance (Grath et al., 2018), and medicine (Caruana et al., 2015). It would be prudent to discourage algorithms that generate misleading explanations. However, since explanations are task, model, and context-specific (Doshi-Velez and Kim, 2017), identifying unfavorable explainability methods proves difficult in practice.

Previous work claims Additive Explainable AI (XAI) methods[1] are harmful if their generated rankings of input importance do not correlate with baseline methods (Jain and Wallace, 2019). Therefore, because tokens ranked by attention weights do not *agree* with two older XAI methods, Jain and Wallace (2019) strengthen their claim that 'attention is not explanation'. Their conclusion is concerning since the attention mechanism (Bahdanau et al., 2015) provides plausible insight into what tokens the model considers relevant for a prediction (Galassi et al., 2020) and is faithful to the underlying computation, providing benefits even when omitted during inference (Pruthi et al., 2020). Thus, any attempt to disqualify a potential explanation should be carefully tested.

Expecting agreement between XAI methods assumes the existence of an ideal or correct explanation. However, it is possible that equally valid, albeit poorly correlated, importance rankings may exist. We claim it is unrealistic to expect XAI methods based on different algorithms to compress a model's complex decision process in the same way. In this work, we hold a selection of more recent XAI methods to the same standard as attention-based explanations to investigate what agreement as an evaluation measure can lead us to conclude. We ask the following research question:

RQ: *How well do the XAI methods LIME, Integrated Gradients, DeepLIFT, Grad-SHAP, and Deep-SHAP correlate **(i)** with each other and **(ii)** with attention-based explanations? Does the correlation depend on **(a)** the model architecture (LSTM- and Transformer-based), or **(b)** the nature of the classification task (single- and pair-sequence)?*

We observe low overall agreement between methods, particularly for a Transformer-based model, and use this empirical evidence along with our theoretical objections to claim that — without making tenuous assumptions — the (lack of) correlation between feature importance rankings is uninformative.

[1] For the sake of brevity, we refer to all feature-additive algorithms (e.g., Ribeiro et al., 2016) simply as 'XAI methods'.

2 Related Work

Jain and Wallace (2019) were the first to compare the *agreement* of attention-based explanations with simple XAI methods. Specifically, they report a weak Kendall-τ correlation between the rankings of input token importance obtained from attention weights and those obtained from the input $\times$ gradient (Kindermans et al., 2016; Hechtlinger, 2016) and leave-one-out (Li et al., 2016) XAI methods. To obtain these rankings, they apply a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) model with an additive ($tanh$) (Bahdanau et al., 2015) attention mechanism to both single- and pair-sequence classification tasks. We test the generalizability of their conclusions by including a more complex Transformer-based model and by comparing more recent XAI methods.

Despite algorithmic concerns with Jain and Wallace (2019)'s approach (Wiegreffe and Pinter, 2019; Grimsley et al., 2020), their influential critique has inspired efforts to enhance the faithfulness and plausibility of attention-based explanations. Proposed modifications of the attention mechanism include guided training (Zhong et al., 2019), sparsity (Correia et al., 2019), the minimization of hidden state conicity (Mohankumar et al., 2020), or the introduction of a word-level objective for recurrent architectures (Tutek and Snajder, 2020). Another strategy addresses problems with analyzing attention weights in their raw form, either by projecting from the null space of multi-head self-attention (Brunner et al., 2020), addressing token identifiability in Transformers (Abnar and Zuidema, 2020), or by accounting for the transformed vectors' magnitude (Kobayashi et al., 2020). In some of these papers, an increased agreement with a small set of XAI methods serves as evidence for an improvement in the attention mechanism's explainability. In contrast, we perform a more extensive comparison with five newer XAI methods.

Complementary to our work, Atanasova et al. (2020) propose a series of diagnostic tests to evaluate XAI methods for text classification. We similarly compare XAI methods, but prescriptions of their desirable properties are outside our scope.

3 Method

We define an *explanation* of an input sequence of tokens as a vector of corresponding importance scores. We investigate two types of explanations: (i) those from recent XAI methods and (ii) those based on attention scores. We measure *agreement* between these explanation methods as the Kendall-τ correlation between the ranked importance scores of all input tokens.

3.1 Recent XAI methods

We select a number of recent XAI methods, namely: LIME (Ribeiro et al., 2016); Integrated Gradients (Sundararajan et al., 2017); DeepLIFT (Shrikumar et al., 2017); and two methods from the SHAP (Lundberg and Lee, 2017) family: Grad-SHAP, which is based on Integrated Gradients; and Deep-SHAP, which is based on DeepLIFT. We do not compare XAI methods to their SHAP approximations. Their agreement is biased due to their algorithmic similarity.

3.2 Attention-based explanations

Given an input sequence of tokens $S = t_1, ..., t_n$, we define an *attention-based explanation* as an assignment of attention weights $\boldsymbol{\alpha} \in \mathbb{R}^n$ over the tokens in S. Since the dimensionality of $\boldsymbol{\alpha}$ is architecture-dependent, it may be necessary to filter or aggregate the weights. In our experiments, this is only relevant for DistilBERT's self-attention mechanism (Vaswani et al., 2017). Previous analyses at the attention head level (e.g., Baan et al., 2019; Clark et al., 2019) implicitly assume that contextual word embeddings remain tied to their corresponding tokens across self-attention layers. This assumption may not hold in Transformers, since information mixes across layers (Brunner et al., 2020). Therefore, we use the *attention rollout* (Abnar and Zuidema, 2020) method — which assumes the identities of tokens are linearly combined through the self-attention layers based exclusively on attention weights — to calculate a post-hoc, faithful token-level attribution. Like Abnar and Zuidema (2020), we use the attribution calculated for the last layer's [CLS] token, resulting in a final vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ at the time of evaluation.

Recurrent models similarly suffer from issues of identifiability. In LSTM-based models, attention is computed over hidden representations across timesteps, which does not provide faithful token-level attribution. Approaches that trace explanations back to individual timesteps (Bento et al., 2020) or input tokens (Tutek and Snajder, 2020) are only just emerging. Therefore, we limit ourselves to an analysis of the raw attention weights.

4 Experiments

4.1 Datasets

We evaluate two types of classification tasks: (i) single-sequence, and (ii) pair-sequence. For single-sequence, we perform binary sentiment classification on the popular Stanford Sentiment Treebank (**SST-2**) (Socher et al., 2013) and the **IMDb** Large Movie Reviews Corpus (Maas et al., 2011). To compare our results with Jain and Wallace (2019), we use identical splits and pre-processing. We also remove sequences longer than 240 tokens to increase inference speed during attribution calculation. For pair-sequence, we examine natural language inference and understanding with the **SNLI** (Bowman et al., 2015), **MultiNLI** (Williams et al., 2018), and **Quora** Question Pairs datasets. Since MultiNLI has no publicly available test set, we use the English subset of the XNLI (Conneau et al., 2018) test set. We use a custom split (80/10/10) for the Quora dataset, removing pairs with a combined count of 200 or more tokens. Readers may refer to our codebase for further details. Most importantly, we contextualize the attention mechanism's utility for each dataset by comparing against a uniform activation baseline (Wiegreffe and Pinter, 2019).

4.2 LSTM-based Model

For our LSTM-based model, we use the same single-layered bidirectional encoder with additive $(tanh)$ attention and a linear feedforward decoder as used by Jain and Wallace (2019). In pair-sequence tasks, we embed, encode, and induce attention over each sequence separately. The decoder predicts the appropriate label from the concatenation of: both context vectors c_1 and c_2; their absolute difference $|c_1 - c_2|$; and their element-wise product $c_1 \cdot c_2$.

4.3 Transformer-based Model

To reduce the computational overhead, we fine-tune the lighter, pre-trained DistilBERT variant (Sanh et al., 2019) instead of the full BERT model (Devlin et al., 2019). For classification, we add a linear layer on top of the pooled output. In pair-sequence tasks, we concatenate sequences with a [SEP] token.

4.4 Training the models

We train three independently-seeded instances of both models using the AllenNLP framework (Gardner et al., 2018), each for a maximum of 40 epochs.

	BiLSTM		DistilBERT	
	Uniform	Softmax	Uniform	Softmax
MNLI	$.659 \pm .001$	$.667 \pm .004$	$.599 \pm .002$	$.779 \pm .002$
Quora	$.829 \pm .001$	$.830 \pm .001$	$.832 \pm .001$	$.888 \pm .001$
SNLI	$.804 \pm .004$	$.807 \pm .002$	$.770 \pm .005$	$.871 \pm .001$
IMDb	$.874 \pm .011$	$.872 \pm .014$	$.879 \pm .003$	$.890 \pm .005$
SST-2	$.823 \pm .008$	$.826 \pm .011$	$.823 \pm .004$	$.842 \pm .003$

Table 1: Test set accuracy when using softmax or uniform activations in the attention mechanisms. A uniform activation renders the mechanism defunct and contextualizes its utility for a particular task.

We use a patience value of 5 epochs for early stopping. For the BiLSTM, we follow Jain and Wallace (2019) and select a 128-dimensional encoder hidden state with a 300-dimensional embedding layer. We tune pre-trained FastText embeddings (Bojanowski et al., 2017) and optimize with the AMSGrad variant (Tran and Phong, 2019) of Adam (Kingma and Ba, 2015). For DistilBERT, we fine-tune the standard 'base-uncased' weights available in the HuggingFace library (Wolf et al., 2019) with the AdamW (Loshchilov and Hutter, 2019) optimizer. Table 1 confirms our models are sufficiently accurate for our analysis. Our extendable Python package for evaluating agreement between XAI methods and attention-based explanations `court-of-xai` is publicly available[2].

4.5 Explaining the models

We leverage existing implementations of Integrated Gradients, DeepLIFT, Grad-SHAP, and Deep-SHAP[3] and use the padding token as a baseline where applicable. Due to resource constraints, we limit the number of samples for each sequence to 250 in our implementation of LIME. We also restrict our calculations of all XAI methods to 500 random instances taken from the test set of the corresponding dataset.

5 Results

5.1 XAI methods do not correlate well

Table 2 displays the Kendall-τ correlations for: (a) the BiLSTM model, and (b) the DistilBERT model. We answer **RQ(i)** and **RQ(ii)** by showing our XAI methods neither agree with each other (mean=0.2229) nor with attention-based explanations (mean=0.1705) across all models and tasks.

[2] `github.com/sfschouten/court-of-xai`
[3] `github.com/pytorch/captum`

5.2 Correlation is model and task dependent

For **RQ(a)**, the agreement between the recent XAI methods is much lower for the DistilBERT model (mean=0.0928) than for the BiLSTM model (mean=0.350). Average agreement between the XAI methods and attention-based explanations is comparable for both models (DistilBERT=0.1560, BiLSTM=0.1849). Regarding **RQ(b)**, the total agreement across all methods is more pronounced for the single-sequence datasets (combined model average=0.2415) than for the pair-sequence datasets (combined model average=0.1728). This difference is particularly noticeable for the agreement between XAI methods applied to the BiLSTM (single-sequence=0.5396, pair-sequence=0.2285).

		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP	
Attn	MNLI	.2391	.2523	.2549	.2473	.2370	
	Quora	.0888	.0143	.0894	.0182	.1017	
	SNLI	.3047	.2566	.3158	.2517	.2938	.1849
	IMDb	.1031	.2188	.2494	.2209	.2309	
	SST-2	.1369	.1093	.1372	.1101	.1400	
LIME	MNLI		.2535	.2176	.2488	.1923	
	Quora		.1064	.1633	.1117	.1357	
	SNLI		.2232	.1790	.2156	.1646	
	IMDb		.2514	.2397	.2505	.2326	
	SST-2		.6230	.5921	.6228	.5538	**.3530**
Int-Grad	MNLI			.4984		.4015	
	Quora			.2906		.2433	
	SNLI			.2461	Grad-SHAP	.2219	
	IMDb			.7331		.7021	
	SST-2			.8683		.8056	

(a) BiLSTM

		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP	
Attn Roll	MNLI	.1595	.1891	.2432	.1905	.2067	
	Quora	.0992	.0574	.2267	.0518	.2257	
	SNLI	.1048	.1645	.2214	.1600	.1796	**.1560**
	IMDb	.0991	.1818	.2516	.1432	.2303	
	SST-2	.1271	.0511	.1328	.0737	.1291	
LIME	MNLI		.1037	.1228	.0969	.1078	
	Quora		.0512	.0809	.0394	.0699	
	SNLI		.0598	.1351	.0550	.0969	
	IMDb		.0775	.0596	.0707	.0558	
	SST-2		.1869	.0726	.1571	.0534	.0928
Int-Grad	MNLI			.2153		.1752	
	Quora			.0625		.0535	
	SNLI			.0955	Grad-SHAP	.0851	
	IMDb			.1433		.1093	
	SST-2			.0498		.0419	

(b) DistilBERT

Table 2: Mean Kendall-τ between the explanations given by our XAI methods for each model when applied to 500 instances of the test portion of each dataset.

6 Discussion & Conclusion

We observe low overall agreement between XAI methods. Since we find XAI methods are prone to disagreement, we believe different methods can yield different inferences about the same model.

Our observation that more complex models and tasks show lower agreement, with some exceptions for the BiLSTM model, may lead us to one of two possible conclusions. If we assume an ideal explanation exists — and that the desirability of an arbitrary XAI method decreases monotonically with correlation (in our case, Kendall's-τ) — then the low agreement we observe means at most one of our selected methods is desirable. Alternatively, we can reject this assumption. In that case, it is difficult to draw conclusions when there is low agreement among XAI methods. Perhaps rankings of model inputs can capture only a narrow slice of the model's behavior such that many equally valid compressions exist. Thus, XAI methods may be in disagreement while remaining faithful.

We observe higher agreement among XAI methods on the simpler models and tasks, but is it possible they are just harmonious in their error? It is unlikely this is the case, assuming that most rankings are undesirable. XAI methods, whose algorithms are (mostly) unrelated, are more likely to agree when selecting from the subset of desirable rankings. However, suppose desirable explanations are common because many faithful rankings exist (as argued above) or because the task is too complicated for humans to judge token-level importance. In that case, we may conclude nothing from higher measures of agreement. For more grounded evaluations of plausibility, agreement with human judgments like those available in e-SNLI (Camburu et al., 2018) may be more informative.

We recommend practitioners investigate the agreement of as many XAI methods as possible to judge each method's utility when applied to a model and task. We will study the agreement between the top-k features selected by XAI methods in future work since they may be of the most interest to the end-user. Top-k token comparison to human-annotated rationales is common (DeYoung et al., 2020; Atanasova et al., 2020) and has demonstrated success when applied to attention-based explanations (Treviso and Martins, 2020). We will also examine more expressive explanation methods, such as those capable of explaining pairwise feature interactions (e.g., Janizek et al., 2020).

Acknowledgments

We thank Maarten de Rijke and the anonymous reviewers for their comments and suggestions. This research was supported by the Dutch National Police, and the Netherlands Organisation for Scientific Research (NWO) under project nr. 652.001.003. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

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.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Understanding multi-head attention in abstractive summarization. *CoRR*, abs/1911.03898.

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*.

J. Bento, Pedro Saleiro, A. F. Cruz, Mário A. T. Figueiredo, and P. Bizarro. 2020. Timeshap: Explaining recurrent models through sequence perturbations. *ArXiv*, abs/2012.00073.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31, pages 9539–9549. Curran Associates, Inc.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA. Association for Computing Machinery.

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 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating crosslingual sentence representations.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. 2019. Adaptively sparse transformers. 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 2174–2184, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2020. Attention in natural language processing. *IEEE*

Transactions on Neural Networks and Learning Systems, page 1–18.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640.

Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. 2018. Interpretable credit application predictions with counterfactual explanations.

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.

Yotam Hechtlinger. 2016. Interpretation of prediction models using the input gradient. *CoRR*, abs/1611.07634.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Alon Jacovi and Yoav Goldberg. 2020. 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.

D. Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing.

Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. 2016. Investigating the influence of noise and distractors on the interpretation of neural networks. *CoRR*, abs/1611.07270.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. 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 4765–4774. Curran Associates, Inc.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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.

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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Richard Socher, John Bauer, Christopher D. Manning, and Andrew Y. Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria. 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 - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Phuong Thi Tran and Le Trieu Phong. 2019. On the convergence proof of amsgrad and a new version. *IEEE Access*, 7:61706–61716.

Marcos Treviso and André F. T. Martins. 2020. The explanation game: Towards prediction explainability through sparse communication. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 107–118, Online. Association for Computational Linguistics.

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.

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 *Advances in neural information processing systems*, pages 5998–6008.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Ruiqi Zhong, Steven Shao, and Kathleen R. McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *CoRR*, abs/1908.06870.