Order in the Court: Explainable AI Methods Prone to Disagreement

Michael Neely [*] [1] Stefan F. Schouten [*] [1] Maurits J. R. Bleeker [1] Ana Lucic [1]

Abstract

By computing the rank correlation between attention weights and feature-additive explanation methods, previous analyses either invalidate or support the role of attention-based explanations as a faithful and plausible measure of salience. To investigate whether this approach is appropriate, we compare LIME, Integrated Gradients, DeepLIFT, Grad-SHAP, Deep-SHAP, and attention-based explanations, applied to two neural architectures trained on single- and pair-sequence language tasks. In most cases, we find that none of our chosen methods agree. Based on our empirical observations and theoretical objections, we conclude that rank correlation does not measure the quality of feature-additive methods. Practitioners should instead use the numerous and rigorous diagnostic methods proposed by the community.

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 & Goldberg, 2020). The rest are irrelevant in the best case and harmful in the worst, particularly in critical domains such as law (Kehl & Kessler, 2017), finance (McGrath et al., 2018), and medicine (Caruana et al., 2015). It would therefore be prudent to discourage algorithms that generate misleading explanations. However, it is challenging to identify when Additive Explainable AI (XAI) methods fail without first decomposing the abstract concepts of plausibility and faithfulness into measurable diagnostic properties.

In their critique of attention-based explanations, Jain & Wallace (2019) argue that faithful XAI methods[2] must be highly *agreeable*[3]. That is, their generated rankings of input importance must correlate with other XAI methods. Following Jain & Wallace (2019)'s claim that 'attention is not explanation', several recent papers have presented an increased agreement with a small set of XAI methods as evidence for their proposed method's ability to improve the faithfulness of the attention mechanism. For example, Mohankumar et al. (2020) show that minimizing hidden state conicity in a BiLSTM improves the Pearson correlation of attention weights with Integrated Gradients (Sundararajan et al., 2017) attributions. As the popularity of *agreement as evaluation* grows (Meister et al., 2021; Abnar & Zuidema, 2020, *inter alia*), we believe it is worth investigating the diagnostic capacity of agreement as a metric.

Under the paradigm of *agreement as evaluation*, proposed XAI methods (e.g., attention-based) are compared to one or more established XAI method(s) (e.g., gradient-based). However, can any XAI method act as the standard against which other XAI methods may be graded? Explanations are task-, model-, and context-specific (Doshi-Velez & Kim, 2017), and the performance of XAI methods depends on the particular diagnostic tests considered (DeYoung et al., 2020; Robnik-Šikonja & Bohanec, 2018, *inter alia*). In this work, we examine the agreement of contemporary XAI methods in a more expansive study to investigate what *agreement as evaluation* can lead us to conclude. We ask:

***RQ:** How well do the XAI methods LIME, Integrated Gradients, DeepLIFT, Grad-SHAP, and Deep-SHAP correlate **(i)** with one 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 XAI methods, particularly for the more complex Transformer-based model, and pair-sequence tasks. We use this empirical evidence, along with our theoretical objections, to argue that practitioners should refrain from grading XAI methods based on agreement. Rank correlation is not a method of objective evaluation unless ground-truth rankings are available (e.g., in Yalcin et al., 2021). In all other situations, rigorous diagnostic measures — such as those proposed by Atanasova et al. (2020) — are better suited for this role.

[*]Equal contribution [1]University of Amsterdam. Correspondence to: Michael Neely <michael@mneely.tech>, Stefan F. Schouten <sfschouten@gmail.com>, Maurits J. R. Bleeker <m.j.r.bleeker@uva.nl>, Ana Lucic <a.lucic@uva.nl>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

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

[3] Ethayarajh & Jurafsky (2021) use the term *consistent*.

2. Related Work

2.1. Agreement as Evaluation

Jain & Wallace (2019) introduced the *agreement as evaluation* paradigm by comparing 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 from the input × gradient (Kindermans et al., 2016; Hechtlinger, 2016) and leave-one-out (Li et al., 2016) XAI methods. Their work inspired others to measure agreement, including Abnar & Zuidema (2020), who demonstrate their *attention-flow* algorithm improves the SpearmanR correlation with the feature-ablation (blank-out) XAI method, and Meister et al. (2021), who show that — under the same experimental setup as Jain & Wallace (2019) — inducing sparsity in the attention distribution decreases agreement with XAI methods. We test the generalizability of agreement as a metric by including a more complex Transformer-based model and by comparing more recent XAI methods.

2.2. Attention as Explanation

Despite concerns with Jain & Wallace (2019)'s approach (Wiegreffe & Pinter, 2019; Grimsley et al., 2020, *inter alia*), their influential critique has inspired enhancements of 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), and word-level objectives (Tutek & Snajder, 2020). Additionally, techniques such as projecting from the null space of multi-head self-attention (Brunner et al., 2020), or accounting for the transformed vectors' magnitude (Kobayashi et al., 2020), address problems with analyzing attention weights in their raw form. Bastings & Filippova (2020) question why the community is concerned with the faithfulness of attention when salience measures already exist. We contribute to the 'attention/explanation' argument by including attention-based explanations in our survey, but do not seek to justify their use.

2.3. Limitations of XAI methods

XAI methods are known to suffer from limitations. For example, Camburu et al. (2019) show that LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) tend to select a token with zero contribution as the most relevant feature, Kindermans et al. (2019) show that saliency methods are not invariant to consistent transformations of model inputs, and Hooker et al. (2019) demonstrate that gradient-based XAI methods are no better than random rankings of importance under their remove-and-retrain approach. Finally, Yalcin et al. (2021) prove the performance of TreeSHAP is inversely correlated with dataset complexity when ground-truth rankings of feature importance are known. Atanasova et al. (2020) unify various evaluation paradigms with a series of diagnostic tests to evaluate XAI methods for text classification. We also compare XAI methods, but only to investigate the suitability of *agreement as evaluation*.

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; Integrated Gradients; DeepLIFT (Shrikumar et al., 2017); and two methods from the SHAP family: Grad-SHAP, which is based on Integrated Gradients; and Deep-SHAP, which is based on DeepLIFT.

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 our Transformer-based model'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 & 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 & 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 & Snajder, 2020) are only just emerging. Therefore, we limit ourselves to an analysis of the raw attention weights for our LSTM-based model.

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 Stanford Sentiment Treebank (**SST-2**) (Socher et al., 2013) and the **IMDb** Large Movie Reviews Corpus (Maas et al., 2011). We use identical splits and pre-processing as Jain & Wallace (2019), but also remove sequences longer than 240 tokens for faster 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. Most importantly, we include a uniform activation baseline to contextualize the attention mechanism's utility (Wiegreffe & Pinter, 2019).

4.2. LSTM-based Model

We use the same single-layered bidirectional encoder with additive $(tanh)$ attention and linear feedforward decoder as Jain & Wallace (2019). In pair-sequence tasks, we embed, encode, and induce attention over each sequence separately. The decoder predicts the 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. We concatenate pair-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. We use a patience value of 5 epochs for early stopping. For the BiLSTM, we follow Jain & 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 & Phong, 2019) of Adam (Kingma & 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 & 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[4].

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

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

4.5. Explaining the models

We leverage existing implementations of LIME, Integrated Gradients, DeepLIFT, Grad-SHAP, and Deep-SHAP[5], and use the padding token as a baseline where applicable. For LIME, we mask tokens as features and use 1000 samples to train the interpretable models. We apply our XAI methods to 500 random instances taken from each test set.

5. Results

5.1. XAI methods rarely correlate with one another

Table 2 displays the Kendall-τ correlations for: (a) the BiLSTM model, and (b) the DistilBERT model. Since the agreement between XAI methods and their SHAP approximations is biased by algorithmic similarity, we do not include their comparisons in our calculations of average agreement. We answer **RQ(i)** and **RQ(ii)** by showing our XAI methods neither agree with each other (mean $= 0.2684$) nor with attention-based explanations (mean $= 0.1736$) across models and tasks.

5.2. Correlation is model and task dependent

For **RQ(a)**, the agreement between non-attention-based XAI methods is lower for DistilBERT (mean $= 0.1088$) than for the BiLSTM (mean $= 0.4281$). Average agreement between the XAI methods and attention-based explanations is comparable for both models (DistilBERT mean $= 0.1658$, BiLSTM mean $= 0.1814$). Regarding **RQ(b)**, the total agreement across all methods is higher for the single-sequence datasets (combined model mean $= 0.273$) than for the pair-sequence datasets (combined model mean $=$ 0.1883). This difference is particularly noticeable for the BiLSTM (single-sequence mean $= 0.4219$, pair-sequence mean $= 0.2308$).

[4] `github.com/sfschouten/court-of-xai`

[5] `github.com/pytorch/captum`

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. Comparisons between methods and their SHAP variants are not representative and thus colored gray.

		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP
Attn	MNLI	.1958	.2523	.2549	.2473	.2370
	Quora	.0363	.0143	.0894	.0182	.1017
	SNLI	.2198	.2566	.3158	.2517	.2938
	IMDb	.2014	.2188	.2494	.2209	.2309
	SST-2	.1326	.1093	.1372	.1101	.1400
LIME	MNLI		.3281	.2444	.3187	.2269
	Quora		.2099	.1900	.2037	.1670
	SNLI		.2673	.1676	.2481	.1566
	IMDb		.6538	.5854	.6486	.5584
	SST-2		.4968	.4734	.4962	.4422
Int-Grad	MNLI			.4984	.8138	.4021
	Quora			.2906	.7420	.2290
	SNLI			.2461	.6535	.2165
	IMDb			.7331	.9409	.6994
	SST-2			.8683	.9707	.8063
DeepLIFT	MNLI				.4987	.6208
	Quora				.3158	.6179
	SNLI				.2557	.5791
	IMDb				.7378	.8593
	SST-2				.8682	.8729
Grad-SHAP	MNLI					.4015
	Quora					.2433
	SNLI					.2219
	IMDb					.7021
	SST-2					.8056

(a) BiLSTM

		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP
Attn Roll	MNLI	.2678	.1891	.2432	.1905	.2067
	Quora	.1622	.0574	.2267	.0518	.2257
	SNLI	.1434	.1645	.2214	.1600	.1796
	IMDb	.1259	.1818	.2516	.1432	.2303
	SST-2	.1359	.0511	.1328	.0737	.1291
LIME	MNLI		.1794	.1526	.1592	.1205
	Quora		.1407	.0032	.1144	.0095
	SNLI		.1529	.0925	.1104	.0593
	IMDb		.1050	.0696	.0929	.0655
	SST-2		.2861	.0618	.2414	.0499
Int-Grad	MNLI			.2153	.4780	.1708
	Quora			.0625	.4674	.0529
	SNLI			.0955	.3932	.0700
	IMDb			.1433	.5495	.1246
	SST-2			.0498	.4987	.0381
DeepLIFT	MNLI				.2324	.4985
	Quora				.0637	.5951
	SNLI				.1181	.5554
	IMDb				.1306	.4830
	SST-2				.0522	.4514
Grad-SHAP	MNLI					.1752
	Quora					.0535
	SNLI					.0851
	IMDb					.1093
	SST-2					.0419

(b) DistilBERT

6. Discussion & Conclusion

The *agreement as evaluation* paradigm assumes — at least implicitly — the desirability of an XAI method decreases monotonically with its correlation to some unobserved 'ideal'. However, there are reasons to doubt whether this assumption holds. For instance, input rankings may only capture a narrow slice of the model's behavior such that many equally faithful compressions exist. And, since many tasks may be too complex for humans to judge token-level importance, there may also be many plausible rankings. While a handful of highly polar tokens are generally indicative of the class label in binary sentiment classification (Sun & Lu, 2020), annotators may be unsure how to rank the other tokens. The problem only gets worse in the pair-sequence setting. For example, if two words indicate a contradiction, which one is more important? There is a reason that rationale collections are normally limited to binary relevance labels or free-form explanations[6]. Thus, when agreement is measured in the presence of multiple faithful and plausible rankings, XAI methods will look deceptively problematic.

We observe low agreement among XAI methods when explaining more complex models and tasks. If we embraced *agreement as evaluation*, we would be obligated to conclude at most one of our chosen XAI methods is near the ideal; implying the other methods cannot explain the more complex Transformer-based model and pair-sequence tasks. Instead, we interpret our results as evidence against the underlying assumptions of *agreement as evaluation*, and conclude that agreement is not a suitable method of evaluation.

Without an external ground-truth explanation (like those constructed by Yalcin et al., 2021), all rank correlation tells us is whether or not two rankings are similar. For this reason, we recommend practitioners stop using *agreement as evaluation*. Instead, we recommend using robust, theoretically-motivated measures of an XAI method's quality, such as those proposed by Atanasova et al. (2020).

Agreement can still be informative, even if it is unsuitable as an evaluation measure. For example, it may reveal how theoretical properties manifest in practice. While algorithms that approximate Shapley Values are normally referenced with the umbrella term 'SHAP', Ethayarajh & Jurafsky (2021) show that *attention flow* (Abnar & Zuidema, 2020) is also a Shapley Value explanation. Interestingly, we observe low agreement between Grad-SHAP and Deep-SHAP (combined model mean $= 0.2839$) and between *attention flow* and our chosen SHAP approximations as well (mean $= 0.1726$, see our supplementary material). As previously argued, this does not mean these methods are wrong, merely that we cannot assume they are interchangeable.

[6] See (Wiegreffe & Marasovic, 2021) and (DeYoung et al., 2020) for good reviews of explainability datasets in NLP.

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

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://www.aclweb.org/anthology/2020.acl-main.385.

Atanasova, P., Simonsen, J. G., Lioma, C., and Augenstein, I. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3256–3274, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.263. URL https://www.aclweb.org/anthology/2020.emnlp-main.263.

Baan, J., ter Hoeve, M., van der Wees, M., Schuth, A., and de Rijke, M. Understanding multi-head attention in abstractive summarization. *CoRR*, abs/1911.03898, 2019. URL http://arxiv.org/abs/1911.03898.

Bastings, J. and Filippova, K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–155, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL https://www.aclweb.org/anthology/2020.blackboxnlp-1.14.

Bento, J., Saleiro, P., Cruz, A. F., Figueiredo, M. A. T., and Bizarro, P. Timeshap: Explaining recurrent models through sequence perturbations. *CoRR*, abs/2012.00073, 2020. URL https://arxiv.org/abs/2012.00073.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://www.aclweb.org/anthology/Q17-1010.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075.

Brunner, G., Liu, Y., Pascual, D., Richter, O., Ciaramita, M., and Wattenhofer, R. On identifiability in transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BJg1f6EFDB.

Camburu, O., Giunchiglia, E., Foerster, J., Lukasiewicz, T., and Blunsom, P. Can I trust the explainer? verifying posthoc explanatory methods. *CoRR*, abs/1910.02065, 2019. URL http://arxiv.org/abs/1910.02065.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. 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, pp. 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2788613. URL https://doi.org/10.1145/2783258.2788613.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. 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*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://www.aclweb.org/anthology/W19-4828.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. XNLI: evaluating cross-lingual sentence representations. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2475–2485. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1269. URL https://doi.org/10.18653/v1/d18-1269.

Correia, G. M., Niculae, V., and Martins, A. F. T. 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)*, pp.

2174–2184, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1223. URL https://www.aclweb.org/anthology/D19-1223.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://www.aclweb.org/anthology/2020.acl-main.408.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning, 2017.

Ethayarajh, K. and Jurafsky, D. Attention flows are shapley value explanations. *CoRR*, abs/2105.14652, 2021. URL https://arxiv.org/abs/2105.14652.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M. E., Schmitz, M., and Zettlemoyer, L. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640, 2018. URL http://arxiv.org/abs/1803.07640.

Grimsley, C., Mayfield, E., and R.S. Bursten, J. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1780–1790, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.220.

Hechtlinger, Y. Interpretation of prediction models using the input gradient. *CoRR*, abs/1611.07634, 2016. URL http://arxiv.org/abs/1611.07634.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Jacovi, A. and Goldberg, Y. 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*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://www.aclweb.org/anthology/2020.acl-main.386.

Jain, S. and Wallace, B. C. 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)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357. URL https://www.aclweb.org/anthology/N19-1357.

Kehl, D. and Kessler, S. A. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. 2017.

Kindermans, P., Schütt, K., Müller, K., and Dähne, S. Investigating the influence of noise and distractors on the interpretation of neural networks. *CoRR*, abs/1611.07270, 2016. URL http://arxiv.org/abs/1611.07270.

Kindermans, P., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un)reliability of saliency methods. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K. (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pp. 267–280. Springer, 2019. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Kobayashi, G., Kuribayashi, T., Yokoi, S., and Inui, K. 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)*, pp. 7057–7075, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/

2020.emnlp-main.574. URL https://www.aclweb.org/anthology/2020.emnlp-main.574.

Li, J., Monroe, W., and Jurafsky, D. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016. URL http://arxiv.org/abs/1612.08220.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lundberg, S. M. and Lee, S. A unified approach to interpreting model predictions. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P11-1015.

McGrath, R., Costabello, L., Van, C. L., Sweeney, P., Kamiab, F., Shen, Z., and Lécué, F. Interpretable credit application predictions with counterfactual explanations. *CoRR*, abs/1811.05245, 2018. URL http://arxiv.org/abs/1811.05245.

Meister, C., Lazov, S., Augenstein, I., and Cotterell, R. Is sparse attention more interpretable? *CoRR*, abs/2106.01087, 2021. URL https://arxiv.org/abs/2106.01087.

Mohankumar, A. K., Nema, P., Narasimhan, S., Khapra, M. M., Srinivasan, B. V., and Ravindran, B. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4206–4216, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.387. URL https://www.aclweb.org/anthology/2020.acl-main.387.

Ribeiro, M. T., Singh, S., and Guestrin, C. ”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, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Robnik-Šikonja, M. and Bohanec, M. *Perturbation-Based Explanations of Prediction Models*, pp. 159–175. Springer International Publishing, Cham, 2018. ISBN 978-3-319-90403-0. doi: 10.1007/978-3-319-90403-0_9. URL https://doi.org/10.1007/978-3-319-90403-0_9.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, pp. 3145–3153. JMLR.org, 2017.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 455–465, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-1045.

Sun, X. and Lu, W. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3418–3428, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.312. URL https://www.aclweb.org/anthology/2020.acl-main.312.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, pp. 3319–3328. JMLR.org, 2017.

Tran, P. T. and Phong, L. T. On the convergence proof of amsgrad and a new version. *IEEE Access*, 7: 61706–61716, 2019. ISSN 2169-3536. doi: 10.1109/access.2019.2916341. URL http://dx.doi.org/10.1109/ACCESS.2019.2916341.

Tutek, M. and Snajder, J. Staying true to your word: (how) can attention become explanation? In *Proceedings of the 5th Workshop on Representation Learning* *for NLP*, pp. 131–142, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.17. URL https://www.aclweb.org/anthology/2020.repl4nlp-1.17.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wiegreffe, S. and Marasovic, A. Teach me to explain: A review of datasets for explainable NLP. *CoRR*, abs/2102.12060, 2021. URL https://arxiv.org/abs/2102.12060.

Wiegreffe, S. and Pinter, Y. 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)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://www.aclweb.org/anthology/D19-1002.

Williams, A., Nangia, N., and Bowman, S. 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)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

Yalcin, O., Fan, X., and Liu, S. Evaluating the correctness of explainable AI algorithms for classification. *CoRR*, abs/2105.09740, 2021. URL https://arxiv.org/abs/2105.09740.

Zhong, R., Shao, S., and McKeown, K. R. Fine-grained sentiment analysis with faithful attention. *CoRR*, abs/1908.06870, 2019. URL http://arxiv.org/abs/1908.06870.

A. Attention Flow

Table 1. Mean Kendall-τ between the explanations given by *attention flow* and our chosen XAI methods for the DistilBERT model when applied to 500 instances of the test portion of each dataset. IMDb is not included among these datasets, because the long sequences made the *attention flow* computation unfeasible.

		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP
Attn Flow	MNLI	.1326	.1251	.2159	.1227	.2148
	Quora	.0853	.2426	.0367	.0241	.2319
	SNLI	.0844	.0753	.2178	.0571	.2149
	SST-2	.1795	.0689	.1286	.0811	.1202

Despite *attention flow*, Grad-SHAP, and Deep-SHAP all (supposedly) being valid Shapley Value explanations, agreement is low.

B. Reproducibility Checklist

In this Appendix, we include information about our experiments from the Reproducibility Checklist.

B.1. For all reported experimental results

B.1.1. A clear description of the mathematical setting, algorithm, and/or model

We clearly explain our methods in Section 3 and our models, datasets, and experiments in Section 4.

B.1.2. Submission of a zip file containing source code, with specification of all dependencies, including external libraries, or a link to such resources (while still anonymized)

Our code is publicly available at `github.com/sfschouten/court-of-xai`

B.1.3. Description of computing infrastructure used

We conducted our experiments on Amazon Web Services `g4dn.xlarge` EC2 instances using an NVIDIA T4 GPU with 16GB of RAM. The version of PyTorch was `1.6.0+cu101`.

B.1.4. Average runtime for each approach

Refer to Table 2 for the average time to train each model on each dataset.

B.1.5. Number of parameters in each model

The DistilBERT model contained 66955779 trainable parameters and the BiLSTM model contained 12553519 trainable parameters, as reported by the AllenNLP library.

B.1.6. Corresponding validation performance for each reported test result

Table 3 details the validation performance of the best model weights for each dataset.

B.1.7. Explanation of evaluation metrics used, with links to code

We evaluate our models by their accuracy. We evaluate the correlation (agreement) between XAI methods using Kendall's-τ. Both of these metrics are explained in Section 3. The code is available at the previously listed URL.

B.2. For all experiments with hyperparameter search

The items in this part of the Reproducibility Checklist are not applicable to our paper.

B.3. For all datasets used

B.3.1. Relevant statistics such as number of examples

Table 4 lists the number of instances in each split of each dataset.

B.3.2. Details of train/validation/test splits

Split details are outlined in Section 4.1. See below for links to each dataset.

B.3.3. Explanation of any data that were excluded, and all pre-processing steps

Details of data exclusion and pre-processing steps are outlined in Section 4.1.

B.3.4. A link to a downloadable version of the data

Links to download versions of all datasets are included in our code repository. For posterity, links to all datasets are listed here: **SST-2**[1], **IMDb**[2], **SNLI**[3], **MNLI**[4], XNLI[5]. Our **Quora** Question Pair dataset will be made available upon publication.

[1] `https://github.com/successar/AttentionExplanation/tree/master/preprocess/SST`

[2] `https://github.com/successar/AttentionExplanation/tree/master/preprocess/IMDB`

[3] `https://nlp.stanford.edu/projects/snli/`

[4] `https://cims.nyu.edu/~sbowman/multinli/`

[5] `https://cims.nyu.edu/~sbowman/xnli/`

	BiLSTM	DistilBERT
MNLI	8.65 ± 0.635	296.228 ± 48.859
Quora	7.567 ± 1.404	380.056 ± 124.911
SNLI	31.495 ± 5.618	126.395 ± 22.909
IMDb	1.122 ± 0.107	24.2 ± 1.212
SST-2	0.216 ± 0.029	2.833 ± 0.65

Table 2. Number of minutes (average ± standard deviation) required to train each model on each dataset reported across three seeds.

	BiLSTM	DistilBERT
MNLI	67.088 ± 0.190	77.338 ± 0.251
Quora	83.232 ± 0.139	88.801 ± 0.055
SNLI	81.535 ± 0.041	87.679 ± 0.075
IMDb	87.975 ± 1.375	88.587 ± 0.489
SST-2	80.696 ± 0.403	83.066 ± 0.692

Table 3. Validation accuracy (average ± standard deviation) of the selected model epoch reported across three seeds.

B.3.5. For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control

We did not collect new data for this paper.

	Training	Validation	Test
MNLI	392702	10000	5000
Quora	323426	40429	40431
SNLI	550152	10000	10000
IMDb	17212	4304	4363
SST-2	8544	1101	2210

Table 4. Number of instances in each split of each dataset before any exclusions based on length (see Section 4.1). Since MultiNLI has no publicly available test set, we use the English subset of the XNLI dataset.