Order in the Court: Explainable AI Methods Prone to Disagreement

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

University of Amsterdam

University of Amsterdam

Agreement as Evaluation

Idea: "Attention weights should correlate (agree) with other feature-additive Explainable AI(XAI) methods." [6]

- Has been used to invalidate Attention as an XAI method. [6, 8]
- Has also been used to justify alternative analyses of the Attention mechanism. [9, 1]

Is this paradigm valid?

- Can XAI methods act as standards for their peers?
- Explanations are context-specific. [4]
- The 'quality' of XAI methods varies wildly depending on the diagnostic task. [11, 3]
- **No one has held non-Attention-based XAI methods to the same standard!**

Example: Measuring Agreement

Task: Movie Review

Explanation Methods: LIME, Feature Ablation, Attention

	An	ungainly	,	humorless	rush	job	...
	6th	4th	7th	1st	2nd	3rd	5th
	6th	3rd	7th	2nd	1st	4th	5th
α	7th	1st	6th	2nd	5th	4th	3rd

$\tau(\cdot,\cdot) = .80$ (strong correlation)

$\tau(\cdot,\alpha) = .33$ (weak correlation)

$\tau(\cdot,\alpha) = .33$ (weak correlation)

Jain and Wallace, 2019 would say "Attention is not Explanation"

Research Question

*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)?*

Method

- Given input tokens $S = t_1, ..., t_n$ and a model's prediction, produce a vector of scores that denote the **importance** of each token for the model's prediction.
- Treating the scores of each feature-additive method as a ranking, calculate the average agreement between each pair of methods using Kendall's-τ.
- For reproducibility, calculate averages over three randomly sampled subsets of 500 test-set instances.

Experiments

- Models:
  - Recurrent model: **BiLSTM** [5]
  - Transformer model: **DistilBERT** [12]
- Tasks:
  - **Single-sequence**: binary sentiment classification (SST, IMDb)
  - **Pair-sequence**: natural language inference (SNLI, MNLI) and paraphrase detection (Quora)
- Modern XAI methods: **LIME**[10], **Integrated Gradients**[14], **DeepLIFT**[13], **Grad-SHAP**[7], **Deep-SHAP**[7].

Results

		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

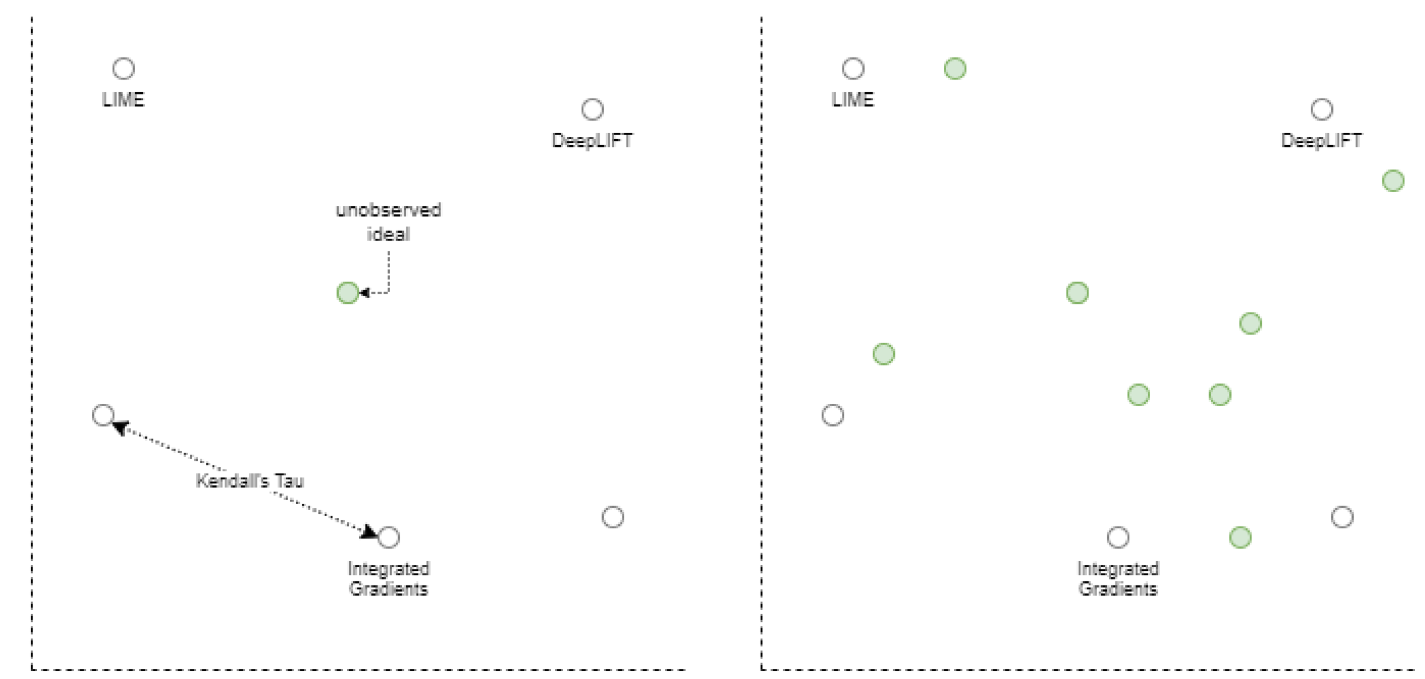
		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP
Attn Roll	MNLI	.2678	.1891	.2432	.1905	.2067
	Quora	.1622	.0574	.2267	.0518	.2257
	SNLI	.1434	.1645	.2214	.1600	.1796
	IMDb	.1259	.1818	.2516	.1432	.2303
	SST-2	.1359	.0511	.1328	.0737	.1291
LIME	MNLI		.1794	.1526	.1592	.1205
	Quora		.1407	.0032	.1144	.0095
	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

Table 1: 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.

Discussion & Conclusion

- The *agreement as evaluation* paradigm assumes a single 'ideal' explanation.



Low average agreement, single ideal explanation. Low average agreement, multiple ideal explanations.

- 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.
- 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 on of our chosen XAI methods is near the ideal.
- Instead, we interpret our results as evidence against the paradigm's underlying assumptions and conclude that *agreement is not evaluation*
- We recommend practitioners instead use theoretically motivated measures of an XAI method's quality[2].

References

[1] Samira Abnar and Willem Zuidema. "Quantifying Attention Flow in Transformers". July 2020. DOI: 10.18653/v1/2020.acl-main.385.

[2] Pepa Atanasova et al. "A Diagnostic Study of Explainability Techniques for Text Classification". Nov. 2020. DOI: 10.18653/v1/2020.emnlp-main.263.

[3] Jay DeYoung et al. "ERASER: A Benchmark to Evaluate Rationalized NLP Models". July 2020. DOI: 10.18653/v1/2020.acl-main.408.

[4] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[5] Alex Graves and Jürgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". 2005. DOI: https://doi.org/10.1016/j.neunet.2005.06.042.

[6] Sarthak Jain and Byron C. Wallace. "Attention is not Explanation". June 2019. DOI: 10.18653/v1/N19-1357.

[7] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[8] Clara Meister et al. "Is Sparse Attention more Interpretable?" 2021. arXiv: 2106.01087.

[9] Akash Kumar Mohankumar et al. "Towards Transparent and Explainable Attention Models". July 2020. DOI: 10.18653/v1/2020.acl-main.387.

[10] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". 2016. DOI: 10.1145/2939672.2939778.

[11] Marko Robnik-Šikonja and Marko Bohanec. "Perturbation-Based Explanations of Prediction Models". 2018. DOI: 10.1007/978-3-319-90403-0_9.

[12] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". 2019. arXiv: 1910.01108.

[13] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning Important Features through Propagating Activation Differences". 2017. URL: http://proceedings.mlr.press/v70/shrikumar17a.html.

[14] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". 2017. URL: http://proceedings.mlr.press/v70/sundararajan17a.html.