Project AI: Do Feature-Additive Explanation Methods Agree?

Stefan F. Schouten

6th June 2021

1 Introduction

The idea central to this Project AI is to structurally compare feature-additive explainability methods. This idea originated when Michael Neely and I were finishing our assignment for the FACT-AI course, and we realized we would like to expand upon the work we were doing. This idea eventually led to us writing a short paper on our findings, which we submitted to a top ML conference.

During FACT-AI our assignment had been to replicate a recent paper on Transparency in AI. The paper we were assigned (Jain and Wallace 2019)[1] has an experiment where attention mechanism are compared to two explainability methods. The explainability methods are designed to produce a score for each input token that can be interpreted as the importance of that token. The question is if the weights given by an attention mechanism are indicating the importance of input tokens as well. They conclude that this is likely not the case in part because: the attention mechanism generates scores that are dissimilar from the two established methods; yet the two explainability methods *do* produce similar scores. The argument is that if attention was an appropriate explainability method that it should have produced similar scores to the established methods. It was this experiment that we chose to focus on in our project.

[1] "Attention is not Explanation" DOI: 10.18653/v1/N19-1357

Because Jain and Wallace open-sourced the code they used for their experiments, merely reproducing their results was not enough. The assignment required us to find additional ways of verifying their claims, thus we wanted to investigate a concern that Jain and Wallace themselves raised:

> "it remains a possibility that agreement is strong between attention weights and feature importance scores for the top-k features only (the trouble would be defining this k and then measuring correlation between non-identical sets)" - Jain and Wallace

To address this we used sparse attention, specifically we defined the top-k as the set of features for which the attention weights were not set to 0 by the sparse activation function, thereby solving the issue of defining k.

After (what at the time seemed like[2]) promising results, we wanted to expand upon this idea by comparing more recent explainability methods. This is what prompted us to contact Ana Lucic, who

[2] Later we found out that the implementation of the top-k rank correlation coefficient we were using was faulty. See Section 2.3 for more details.

had been one of our lecturers during the FACT-AI course. After discussing options, we agreed to perform this expanded experiment in the form of a Project AI to be supervised by Ana Lucic and Maurits Bleeker. This project ultimately resulted in: (1) short paper submissions to two top ML conferences; (2) us presenting the results on two occasions; and (3) a Python package allowing others to perform similar experiments.

This report will focus on detailing my contribution to this project. It will also detail some of the work we did that did not end up in our short paper submission(s). Finally, it will have an expanded explanation of our analysis of the results, which will go into more detail than was possible for the short paper.

2 *Background and Methodology*

In this section some necessary background will briefly be explained. It will also detail the methodology of our experimentation. It will lay out which eXplainable AI methods we compare (2.1), the model architectures we try to explain (2.2), and how we compare the generated explanations (2.3). Finally, we give some terminology and background on the evaluation of XAI methods (2.4).

2.1 *XAI (eXplainable AI) Methods*

XAI Methods attempt to explain the workings of models that are otherwise hard to interpret. Many kinds of these methods exist, but as the title of the project suggests we focus mainly on the group of methods labelled 'feature-additive'. This class of XAI methods attempts to explain a model's output in terms of its inputs. Specifically, they assign an importance weight to each of the model's input features. They do this in an 'additive' fashion, meaning that they assume responsibility for the output can be linearly attributed to the inputs. Each input feature's importance is independent of the importance of the other input features.

Of these feature-additive XAI methods we chose the following five for our experiments: Integrated Gradients, LIME, DeepLIFT, GradSHAP and DeepSHAP.[3]

[3] This choice of methods was partially informed by which methods had an implementation that easily integrated into the implementation of our experiments, also see Section 3.2.

2.2 *Models*

For a description of our models we will simply refer to the description in the ACL paper submission, which can be seen in Figure 1.

For both models we also had two variants, which did not make it into the short paper submission to ACL. These model variants are described below.

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.

Figure 1: Model descriptions from the ACL submission (see Appendix B)

Variant: Sparse Attention Mechanisms. From the start of the project we were interested in the influence of sparsity on explainability. So we planned to have a variant of both models where the attention mechanism would be modified such that instead of a softmax, it used a sparse activation function.

Inspired by recent work (Correia et al. 2019)[4] we particularly wanted to include the α-Entmax activation function. The α-Entmax family of activation functions (Peters et al. 2019)[5] has both the softmax and sparsemax as special cases, with $\alpha = 1$ and $\alpha = 2$ respectively. These functions can be used in place of the Softmax activation function commonly used in attention mechanisms to create sparse attention mechanisms. This can even be done in such a way that the α is a learnable model parameter, creating an adaptively sparse model (Correia et al. 2019).

[4] "Adaptively Sparse Transformers" DOI: 10.18653/v1/D19-1223

[5] "Sparse Sequence-to-Sequence Models" DOI: 10.18653/v1/P19-1146

Variant: Attention Vector Magnitudes instead of Attention Weights. In (Kobayashi et al. 2020)[6], an attempt is made to improve the explanatory value of attention mechanisms. The authors argue that for some attention mechanisms (like those used in a Transformer) it is not the attention weight alone that is indicative of importance. For attention mechanisms like exists in the Transformer, the input vectors are first projected to key, query, and value spaces. The output of the mechanism is a linear combination of the value vectors, so the final contribution that each input has in the output is not just determined by the attention weights but also by the magnitude of the value vectors. Therefore, in this variant the importance scores are obtained as $\alpha_i \cdot |v_i|$ where α_i and v_i are the attention weight and value vector of the i-th token respectively.

[6] "Attention is Not Only a Weight: Analyzing Transformers with Vector Norms" DOI: 10.18653/v1/2020.emnlp-main.574

2.3 *Comparing Importance Weights*

Our primary method of comparing importance weights is by using the Kendall rank correlation coefficient, also known as Kendall's τ. This was the method used by Jain and Wallace, and as such also our first choice. We recognized though that there is no objectively best correlation coefficient to use here. Different coefficients measure different kinds of similarities. Therefore, we also calculated correlations using Spearman's ρ and Pearson's r, but trends were the same regardless of which we used. Therefore these results are not included or discussed in this report.

Throughout the project we planned to also include experiments that looked only at the top-k most relevant tokens. This turned out to be a lot more challenging than expected. Initially we were optimistic because Jain and Wallace already found a top-k generalization of Kendall's τ (Fagin, Kumar and Sivakumar 2003)[7], which they had included in their codebase[8] that could seemingly be used 'out-of-the-box'. However at some point we started doubting the correctness of this implementation, and upon further inspection we became more and more convinced it was incorrect. I opened an issue on Github[9] asking some clarifying questions, but they unfortunately remained unanswered.

We decided to implement the algorithm described in (Fagin, Kumar and Sivakumar 2003) ourselves instead. Although we succeeded, we quickly realized that this particular top-k generalization did not properly take into account ties in the ranking. We noticed that the XAI methods that occasionally produced the same importance score for multiple input tokens were getting higher correlations overall. In hindsight this makes sense, producing a partial ranking is easier than a full ranking. In fact, when all tokens are given the same score and thus all share rank 1, the correlation given by this top-k generalisation is always 1. However sensible this may be on some situations, this was not the kind of comparison we wanted to make.[10]

Since we did not want to restrict ourselves to XAI methods that always produce full rankings, we started looking for different top-k generalisations of Kendall's τ, specifically one that generalises to partial rankings in an unbiased way. Luckily, a later paper (Fagin, Kumar, Mahdian et al. 2004)[11] extends the top-k generalisation in exactly the way we needed, so we set out to implement this method instead. Unfortunately we never gained full confidence in our implementation, and doubts remained about how to interpret the correlations produced by this method. Therefore these results are not included or discussed in this report.

[7] "Comparing Top k Lists" DOI: 10.1137/S0895480102412856

[8] The original author's repository containing the implementation: https://github.com/AlessandroChecco/top-k-kendall-tau

[9] https://github.com/AlessandroChecco/top-k-kendall-tau/issues/1

[10] Note that the regular Kendall's τ (not top-k) has variants that do not suffer from this problem.

[11] "Comparing and aggregating rankings with ties" DOI: 10.1145/1055558.1055568

2.4 *Evaluating Explanations*

"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)".[12] This opening sentence of our ACL submission introduces the vitally important notions of *plausibility* and *faitfulness*. We call the extent to which an explanation is both *plausible* and *faithful* the explanation's *desirability*. It is important to note that these quantities are very difficult to measure. We certainly do not have access to any 'ground-truth' explanations we can use to quantify the desirability. And even hand-crafted 'gold-rated' explanations are problematic, because the entire reason we are trying to improve the eXplainability of AI is because models are increasingly hard to comprehend. This clearly calls in to question how *faithful* these 'gold-rated' explanations can ever be assumed to be.[13]

[12] From ACL submission, also see Appendix B

[13] 'Gold-rated' explanations could still be used to quantify *plausibility*, since it seems reasonable to assume that explanations created by people are in fact plausible to them.

3 *Implementing the Experiments*

For all of our experiments we use the AllenNLP (Gardner et al. 2018)[14] Python library. This framework allows for easily replicable experimentation, and implementing your methods as part of this framework allows others to re-use those methods for their own experimentation.

Large parts of this work were done collaboratively, making a clear separation of who did what quite difficult. I will mention those cases where I feel like I made a considerably large contribution. For a more detailed account of what exactly my contribution was, the set of pull requests authored by me[15] might be informative.

[14] "AllenNLP: A Deep Semantic Natural Language Processing Platform" DOI: 10.18653/v1/W18-2501

[15] https://github.com/sfschouten/court-of-xai/pulls?q=author%3Asfschouten

3.1 *Datasets*

We investigate interpretability on two types of classification tasks, namely: single-sequence, and pair-sequence. The single-sequence datasets we use are: the Stanford Sentiment Treebank (SST-2) (Socher et al. 2013)[16], and the IMDb Large Moview Reviews Corpus (Maas et al. 2011)[17]. The pair-sequence datasets we use are SNLI (Bowman et al. 2015)[18], MultiNLI (Williams et al. 2018)[19], and the Quora Question Pairs dataset.

Of these datasets, AllenNLP already contained code to load SST-2, SNLI, and MultiNLI without effort. To make the Quora dataset work with our models we needed the class responsible for reading the dataset to be able to concatenate each pair of sequences (rather then provide them separately). To do this I had to modify a class that was part of the AllenNLP library, for which I opened a Pull Request[20].

[16] "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank" URL: https://www.aclweb.org/anthology/D13-1170

[17] "Learning Word Vectors for Sentiment Analysis" URL: https://www.aclweb.org/anthology/P11-1015

[18] "A large annotated corpus for learning natural language inference" DOI: 10.18653/v1/D15-1075

[19] "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference" DOI: 10.18653/v1/N18-1101

[20] https://github.com/allenai/allennlp-models/pull/119

The code to read the IMDb dataset was created by Michael.

3.2 XAI Methods

The AllenNLP library has an interface through which it makes 'interpreters' available, this is what AllenNLP calls the classes that implement XAI methods. Early on in the project I built various additional interpreters that utilize the same interface. Initially I implemented an interpreter for leave-one-out, which was entirely my own implementation. I implemented a wrapper that exposed an existing LIME implementation[21] through this same interface. And in order to utilize the attention weights as estimates of token saliency I also implemented an interpreter that simply outputs those weights.

[21] https://github.com/marcotcr/lime

Later on we found a library that implemented a whole range of XAI methods called Captum (Kokhlikyan et al. 2020)[22]. Similar to my LIME wrapper, I also implemented a wrapper for this Captum library that allowed the methods contained in it to be used within AllenNLP.

[22] "Captum: A unified and generic model interpretability library for PyTorch" URL: http://arxiv.org/abs/2009.07896

3.3 Models

The model architectures we use were not implemented in AllenNLP by default. Each model's implementation can be found in our repository, although the implementation of DistilBERT was mostly borrowed from HuggingFace. Those modifications that were necessary were mostly made by Michael. The code for the LSTM-based model we use for our *single*-sequence tasks is based on our implementation created for the FACT-AI course. The implementation of the LSTM-based *pair*-sequence classifier was created by me, inspired by the code of the *single*-sequence variant.

4 Interpreting the Results

In the end we ran experiments across 5 datasets, for 5+1 (the +1 being attention-based saliency) XAI methods, on 2 model architectures with each an additional sparse variation. Note that we make a measurement of Kendall's τ for each combination of XAI method and we have not even included the two kinds of attention aggregation used for the Transformer-based model, or the norm-based variants of the attention-based saliency. Luckily, we were able to analyze much of these results by beginning with the most important configurations and checking if the same trends hold for the others, rather than analysing each entirely independently.

		BiLSTM		DistilBERT		
Method	Task	$\bar{\tau}_{\text{attn}}$	$\bar{\tau}_{\text{FI}}$	$\bar{\tau}_{\text{attn\_flow}}$	$\bar{\tau}_{\text{attn\_roll}}$	$\bar{\tau}_{\text{FI}}$
LIME	Single	.1200	.4207	.1639	.1131	.0917
	Pair	.2109	.1843	.1098	.1212	.0850
IntGrad	Single	.1640	.6636	.0689	.1165	.1033
	Pair	.1744	.2740	.0791	.1370	.0980
GradSHAP	Single	.1655	.6645	.0811	.1084	.0936
	Pair	.1724	.2792	.0680	.1341	.1021
DeepSHAP	Single	.1854	.6333	.1202	.1797	.0705
	Pair	.2109	.2452	.2205	.2040	.0980
DeepLIFT	Single	.1933	.6732	.1286	.1922	.0847
	Pair	.2200	.2961	.2255	.2305	.1251
Average		.1849	.3904	.1335	.1560	.0959

(a) Softmax / Weight-based

LIME	Single	.2155	.4097	.1393	.0649	.0738
	Pair	.2156	.1705	.1211	.0768	.0905
IntGrad	Single	.2945	.6045	.1001	.1230	.0934
	Pair	.1864	.2448	.1314	.1908	.1146
GradSHAP	Single	.2956	.6058	.0903	.1139	.0834
	Pair	.1845	.2526	.1094	.1693	.1067
DeepSHAP	Single	.3201	.5762	.0961	.1461	.0725
	Pair	.2368	.2264	.2077	.1208	.1042
DeepLIFT	Single	.3425	.6055	.1214	.1759	.0870
	Pair	.2494	.2659	.2174	.1299	.1301
Average		.2462	.3573	.1454	.1324	.0974

(b) Entmax / Weight-based

		$\bar{\tau}_{\lvert\text{attn}\rvert}$	$\bar{\tau}_{\text{FI}}$	$\bar{\tau}_{\lvert\text{attn\_flow}\rvert}$	$\bar{\tau}_{\lvert\text{attn\_roll}\rvert}$	$\bar{\tau}_{\text{FI}}$
LIME	Single	.1356	.4207	.1379	.1171	.0917
	Pair	.2095	.1843	.0995	.1263	.0850
IntGrad	Single	.1828	.6636	.0631	.1152	.1033
	Pair	.1738	.2740	.0741	.1119	.0980
GradSHAP	Single	.1843	.6645	.0688	.1032	.0936
	Pair	.1717	.2792	.0647	.1044	.1021
DeepSHAP	Single	.1967	.6333	.1066	.1651	.0705
	Pair	.2086	.2452	.1994	.2158	.0980
DeepLIFT	Single	.2078	.6732	.1246	.1863	.0847
	Pair	.2171	.2961	.2053	.2379	.1251
Average		.1903	.3904	.1215	.1505	.0959

(c) Softmax / Norm-based

LIME	Single	.2163	.4097	.1391	.0886	.0738
	Pair	.2143	.1705	.1089	.1262	.0905
IntGrad	Single	.2953	.6045	.0854	.1185	.0934
	Pair	.1865	.2448	.1086	.1784	.1146
GradSHAP	Single	.2965	.6058	.0787	.1065	.0834
	Pair	.1851	.2526	.0923	.1548	.1067
DeepSHAP	Single	.3205	.5762	.1027	.1595	.0725
	Pair	.2377	.2264	.1876	.1952	.1042
DeepLIFT	Single	.3430	.6055	.1295	.1953	.0870
	Pair	.2494	.2659	.1944	.2175	.1301
Average		.2465	.3573	.1306	.1581	.0974

(d) Entmax / Norm-based

Table 1: Mean correlation between each feature importance method and (i) variants of attention ($\bar{\tau}_{\text{attn\_...}}$, $\bar{\tau}_{\lvert\text{attn\_...}\rvert}$) and (ii) other feature importance methods ($\bar{\tau}_{\text{FI}}$). Each sub-table displays the correlations for a variant of attention (α-Entmax vs. Softmax activation function and weight-based vs. norm-based saliency). These values are obtained by averaging across three independently seeded runs for each (model, dataset) pair, and then averaging across the datasets for each task. Correlations between feature importance methods and their SHAP equivalents (e.g. Integrated Gradients and Gradient SHAP) are not included in the averages. The task column specifies which rows pertain to agreement on 'Single'-sequence classification vs. 'Pair'-sequence classification.

4.1 The results

In Table 1 we can see the most important results of the project summarized. Specifically these are the results that used Kendall's τ as the correlation metric. Sub-table 1(a) has the results that made it into the paper submissions.

What follows is an overview of the questions we asked, and the answers suggested by the data in Table 1.

1. How well do recent XAI methods agree with one another?
   - ■ Generally speaking recent XAI methods correlate only to a relatively low degree (the overall average is 0.2353).

   (a) Does their agreement depend on model architecture (LSTM- or Transformer-based)?
   - ■ Yes, agreement among XAI methods on the BiLSTM (0.3904, 0.3573) is significantly higher than on the DistilBERT (0.0959, 0.0974).[23]

   (b) Does their agreement depend on the nature of the classification task (single- or pair-sequence)?
   - ■ Yes, for the BiLSTM agreement is significantly higher for the single-sequence datasets (0.6111, 0.5603) than for pair-sequence datasets (0.2558, 0.2320).[24]
   - ■ For the DistilBERT the agreement is overall so low, it is hard to say whether or not there is a real difference.

2. How well does attention-based saliency agree with recent XAI methods?

   (a) Does a sparse attention activation function improve agreement of attention with the chosen XAI methods?
   - ■ For the BiLSTM, yes the agreement between attention and the XAI methods improves from an average correlation of .1849 to .2462 (weight-based) and from .1903 to .2465 (norm-based).
   - ■ For the DistilBERT, no not clearly. The differences are mixed and minimal, showing a slight increase for attention flow (weight- and norm-based); and showing either a slight decrease or a slight increase for attention rollout when using the weight-based or norm-based attention saliency respectively.

   (b) Does the use of norm-based attention saliency (rather than weight-based) improve agreement with the chosen XAI methods?

[23] Values in parentheses are averages for the models with a Softmax and α-Entmax activation function respectively.

[24] See note 23.

■ No, not clearly. The difference for the BiLSTM is a very small positive one. The differences are slightly larger for the DistilBERT but are also more mixed. For one of the attention-variants (attention rollout with Softmax) there is a small negative effect.

(c) Which method of attention aggregation used for the DistilBERT results in the highest agreement?

■ The differences are again small. However, agreement is highest with attention rollout for the softmax/weight-based, softmax/norm-based and entmax/norm-based variants. Only for the entmax/weight-based variant is attention flow higher.

4.2 Interpretation

For quite some time we struggled with how to interpret these results. Specifically with what conclusions may be drawn from the overall low agreement among the establish XAI methods.

Jain and Wallace use attention-based saliency's lack of agreement to argue against it as a form of explanation. If we were to follow this line of reasoning then we would have to conclude that not a single XAI method is currently equipped to explain the behaviour of the DistilBERT model or of either model when applied to the pair-sequence classification task. However, maybe the line of reasoning is not valid. In that case we are effectively saying that agreement is not a good method of evaluation.

Either side of this dichotomy represents a relatively strong claim. Therefore we decided that we ought to properly break down the line of reasoning before deciding which side to accept.

Agreement as evaluation can be broken down into two questions: 1) when (if at all) does low agreement imply low desirability? and 2) when (if at all) does high agreement imply high desirability? This is also how we broke it down in our ACL submission as can be seen in Figure 2.

A useful way to think about agreement is to think of it in terms of a 'space of explanations'. This (discrete[25]) space has at each point a different explanation. If these explanations are rankings of input features, then the distances between them might be given by the Kendall-τ-distance (number of discordant pairs between the rankings).[26] And we can think of desirability as a function defined on this space. This way of thinking about the explanations is what I used when creating the diagrams that can be seen in Figure 3 (see also Section 6). In this Figure we see variants of what the 'space of ex-

[25] It is discrete when what we mean by explanation is a ranking, if we mean a vector of saliency scores it is continuous.

[26] This distance is also used in the Kendall's τ coefficient as follows:

$$\tau = 1 - 2\frac{d}{\binom{n}{2}}$$

where d is the number of discordant pairs, the distance.

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.

Figure 2: Discussion of *agreement as evaluation* from ACL submission.

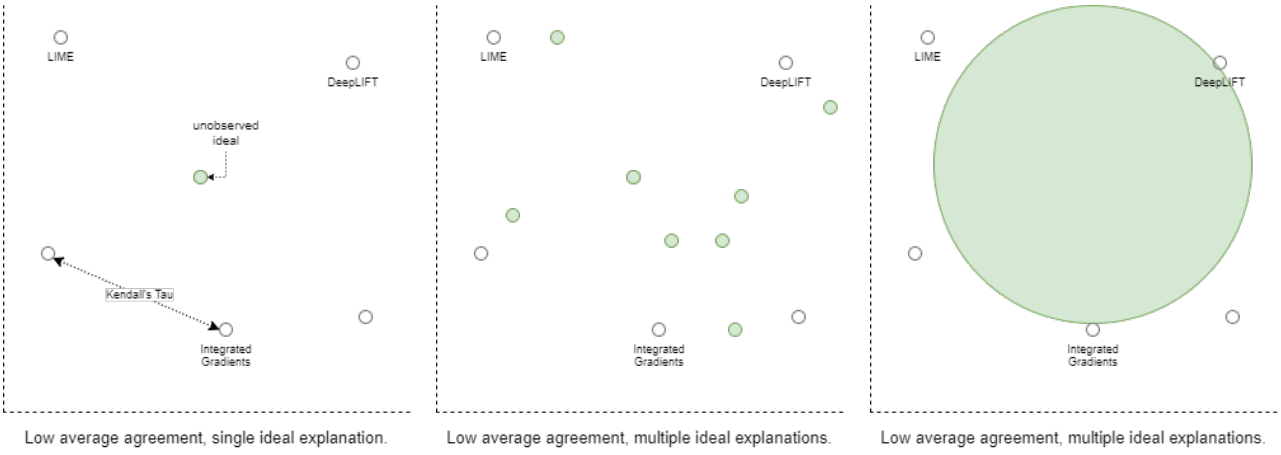


Figure 3: Diagrams of various scenarios depicting low average agreement.

planations' might look like when there is low average agreement. In each of the diagrams depicted the green circles represent peaks in the desirability function.

The leftmost diagram in the figure depicts the most presumptuous interpretation of *agreement as evaluation*. It relies on strong assumptions: "If we assume a [single] ideal explanation exists – and that the desirability of an arbitrary XAI method decreases monotonically with correlation [to this ideal explanation] – then the low agreement we observe means at most one of our selected methods is desirable"[27]. Indeed the leftmost diagram suggests only a single method may be close to the unobserved ideal at a time, but that it is more likely none of them are. Thus, under these assumptions we may say that low agreement means the methods are on average removed far from the ideal, making a claim about the methods' desirability collectively. Making claims about the desirability of a single method is more difficult, even under these assumptions. In fact, the best we can do for a single method is a probabilistic argument. We are more likely to observe low agreement for a single method and no others in a world where the ideal is close the methods with high agreement (assuming no other evidence for or against the XAI methods).

[27] From ACL submission, see Figure 2.

However, if we move away from the strong assumptions, for example by allowing multiple peaks (local maxima) of desirability as seen in the middle diagram. Or by having a multiple ideal explanations (a plateau of desirability) as seen in the rightmost diagram. Then we can no longer conclude that at most one of the methods' explanations can be highly desirable. Without the assumptions we really cannot infer anything from low agreement. And indeed these kind of assumption-breaking examples seem plausible. "Perhaps rankings of model inputs can capture only a narrow slice of the model's behaviour such that many equally [*faithful*] compressions exist"[28]. Or "the task [we wish to explain] is too complicated for humans to judge token-level importance"[29] such that many equally *plausible* rankings exist.

[28] See note 27

[29] See note 27

In Figure 4 we can see a depiction of high average agreement. Here we can see that the explanations produced by the various XAI methods are seemingly clustering around an unobserved ideal. Despite not knowing where the ideal is, we believe that when methods are clustered like this, that it should be close to an unobserved ideal because if not, then *why are they clustered?* "Is it possible they are just harmonious in their error"[30], i.e. biased? This does not seem likely given that the XAI methods' algorithms are (mostly) unrelated. In other words we reject a null hypothesis stating that XAI methods' explanations are distributed randomly in favor of a hypothesis stating that the XAI methods are uncovering something meaningful.

[30] See note 27

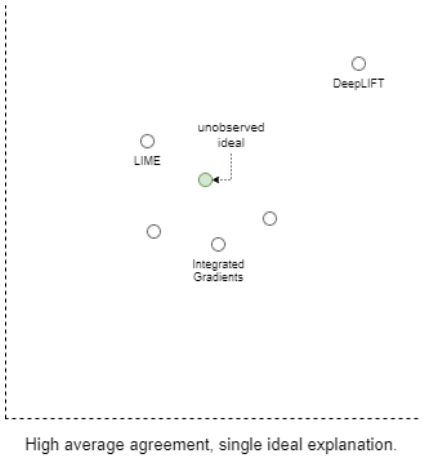


Figure 4: Diagram depicting likely scenario with high average agreement.

This argument works regardless of how we feel about the previously discussed assumptions. [31]

[31] In fact, observing clustering would count as evidence in support of the assumptions, since observing clustering is more likely in a world where the assumptions hold.

5 *Writing the Papers*

One of the goals of this project was always to write a paper for submission at a conference. After we had obtained the bulk of our results, and we were starting to realize how to interpret them, we had to decide what kind of paper to write.

What we saw in our results was not a clear increase in agreement for the sparse and/or norm-based variants of attention as we had hoped. And more importantly we saw a surprising lack of agreement for the DistilBERT model and the pair sequence task that we did not at all expect. At the start of the project we had more or less assumed that the correlation would be high for the chosen XAI methods and merely act as a reference for the agreement of attention based saliency.

In response to this I argued for a short paper focusing entirely on what I considered our most interesting observation, the surprising lack of agreement. And this is the angle we went for when writing our first submission for the EACL conference.

At the time we wrote this submission our understanding of how to interpret the observations had not yet crystallized completely. This caused us to word a few things poorly and the reviewers[32] caught on to this. One reviewer in particular (Review 2 in the Appendix) made a number of good points that caused us to go back to and think more about the interpretation of our observations. The general response to the reviewers is where we first describe the dichotomy that I described at the start of Section 4.2.

[32] See Appendix **??**.

After we were rejected for publication at EACL we decided to try

again. This time we would submit to ACL-IJCNLP and shift the focus more onto the 'agreement as evaluation' and discussion of how to interpret the low agreement, resulting in the discussion shown in Figure 2.

6 *Presenting Our Work*

We were invited to present our work on two occasions. The first was during the 2021 edition of the FACT-AI course. The second was at one of the IRLab's Soos talks. At the time of both of these talks Michael was living in the United States, for him both talks were scheduled for the middle of the night. Thus we agreed that I would present these talks. In exchange, Michael put together the slideshows[33], which I used with some changes[34].

[33] See Appendix E and D.

[34] The biggest of which were the diagrams of Figure 3 and 4.

7 *Conclusion*

In this project we have systematically analyzed the extent to which recent XAI methods agree. We find a serious lack of agreement on more complex models and tasks. Furthermore, we question strongly on theoretical grounds how much can be concluded from (a lack of) rank correlation in the first place. We expect that these results will be helpful to the conversation on how to evaluate XAI methods in the future.

During the project we have not just practiced doing research, writing and presenting; but we have also learned about the process of getting your work published. This has been extremely valuable, and I have enjoyed working on this project immensely.

Acknowledgement

I would like to thank our supervisors, Ana Lucic and Maurits Bleeker for their great advice throughout and their continued encouragement.

References

Bowman, Samuel R. et al. (Sept. 2015). "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2015. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642. DOI: 10.18653/v1/D15-1075. URL: https://www.aclweb.org/anthology/D15-1075 (visited on 03/06/2021).

Correia, Gonçalo M., Vlad Niculae and André F. T. Martins (Nov. 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)*. EMNLP-IJCNLP 2019. Hong Kong, China: Association for Computational Linguistics, pp. 2174–2184. DOI: 10.18653/v1/D19-1223. URL: https://www.aclweb.org/anthology/D19-1223 (visited on 12/05/2021).

Fagin, Ronald, Ravi Kumar, Mohammad Mahdian et al. (14th June 2004). "Comparing and aggregating rankings with ties". In: *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. PODS '04. New York, NY, USA: Association for Computing Machinery, pp. 47–58. ISBN: 978-1-58113-858-0. DOI: 10.1145/1055558.1055568. URL: http://doi.org/10.1145/1055558.1055568 (visited on 30/04/2021).

Fagin, Ronald, Ravi Kumar and D. Sivakumar (1st Jan. 2003). "Comparing Top k Lists". In: *SIAM Journal on Discrete Mathematics* 17.1. Publisher: Society for Industrial and Applied Mathematics, pp. 134–160. ISSN: 0895-4801. DOI: 10.1137/S0895480102412856. URL: https://epubs-siam-org.proxy.uba.uva.nl/doi/10.1137/S0895480102412856 (visited on 30/04/2021).

Gardner, Matt et al. (July 2018). "AllenNLP: A Deep Semantic Natural Language Processing Platform". In: *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1–6. DOI: 10.18653/v1/W18-2501. URL: https://www.aclweb.org/anthology/W18-2501 (visited on 03/06/2021).

Jacovi, Alon and Yoav Goldberg (July 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*. ACL 2020. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://www.aclweb.org/anthology/2020.acl-main.386 (visited on 12/05/2021).

Jain, Sarthak and Byron C. Wallace (June 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)*. NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://www.aclweb.org/anthology/N19-1357 (visited on 16/03/2021).

Kobayashi, Goro et al. (Nov. 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). EMNLP 2020. Online: Association for Computational Linguistics, pp. 7057–7075. DOI: 10.18653/v1/2020.emnlp-main.574. URL: https://www.aclweb.org/anthology/2020.emnlp-main.574 (visited on 03/06/2021).

Kokhlikyan, Narine et al. (16th Sept. 2020). "Captum: A unified and generic model interpretability library for PyTorch". In: *arXiv:2009.07896 [cs, stat]*. arXiv: 2009.07896. URL: http://arxiv.org/abs/2009.07896 (visited on 03/06/2021).

Maas, Andrew L. et al. (June 2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. ACL-HLT 2011. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150. URL: https://www.aclweb.org/anthology/P11-1015 (visited on 03/06/2021).

Peters, Ben, Vlad Niculae and André F. T. Martins (July 2019). "Sparse Sequence-to-Sequence Models". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. ACL 2019. Florence, Italy: Association for Computational Linguistics, pp. 1504–1519. DOI: 10.18653/v1/P19-1146. URL: https://www.aclweb.org/anthology/P19-1146 (visited on 12/05/2021).

Socher, Richard et al. (Oct. 2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2013. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1631–1642. URL: https://www.aclweb.org/anthology/D13-1170 (visited on 03/06/2021).

Williams, Adina, Nikita Nangia and Samuel Bowman (June 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)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122. DOI: 10.18653/v1/N18-1101. URL: https://www.aclweb.org/anthology/N18-1101 (visited on 03/06/2021).

A Full Tables

	BiLSTM			DistilBERT		
	Uniform	Softmax	Entmax	Uniform	Softmax	Entmax
MNLI	.659 ± .001	.667 ± .004	**.671 ± .003**	.599 ± .002	**.779 ± .002**	.763 ± .006
Quora	.829 ± .001	.830 ± .001	**.837 ± .001**	.832 ± .001	**.888 ± .001**	**.888 ± .001**
SNLI	.804 ± .004	.807 ± .002	**.813 ± .002**	.770 ± .005	**.871 ± .001**	.869 ± .002
IMDb	.874 ± .011	.872 ± .014	**.892 ± .003**	.879 ± .003	.890 ± .005	**.890 ± .001**
SST-2	.823 ± .008	.826 ± .011	**.828 ± .004**	.823 ± .004	.842 ± .003	**.844 ± .003**

Table 2: 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. Values are obtained from three independently seeded runs and reported as mean ± standard deviation.

		‖Attn‖	LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP
Attn	MNLI	.9611	.2497	.2472	.2722	.2408	.2545
	Quora	.981	.092	.028	.139	.0351	.1469
	SNLI	.9558	.3051	.2839	.337	.2775	.309
	IMDb	.9825	.1378	.2856	.357	.2874	.3365
	SST-2	.9869	.2932	.3034	.328	.3039	.3037
‖Attn‖	MNLI		.2495	.2503	.2754	.2447	.2583
	Quora		.0911	.0274	.1368	.0342	.1457
	SNLI		.3022	.2818	.3361	.2763	.3091
	IMDb		.1381	.2874	.358	.2891	.3375
	SST-2		.2945	.3032	.3279	.3038	.3035
Lime	MNLI			.2301	.1931	.2259	.1719
	Quora			.0646	.1373	.0777	.1222
	SNLI			.2291	.1988	.2184	.1771
	IMDb			.2372	.218	.2359	.2124
	SST-2			.629	.5722	.6222	.5504
Int-Grad	MNLI				.4465	.7922	.3713
	Quora				.1799	.6478	.155
	SNLI				.2834	.6723	.2435
	IMDb				.6492	.9124	.6183
	SST-2				.767	.9336	.7261
DeepLIFT	MNLI					.4449	.6194
	Quora					.2155	.6204
	SNLI					.294	.5823
	IMDb					.6566	.875
	SST-2					.7702	.8552
Grad-SHAP	MNLI						.3693
	Quora						.1791
	SNLI						.2484
	IMDb						.6246
	SST-2						.7254

Table 3: Mean Kendall-τ correlation between each pair of explanation methods for the **BiLSTM** model with **α-entmax** attention applied to 500 randomly selected instances from the test portion of each dataset. Values are averaged across three independently seeded runs. **Key**: Attn = Attention Weights, ||Attn...|| denotes weighted vector norm analysis.

		||Attn||	LIME	IntGrad	DeepLIFT	GradSHAP	DeepSHAP
Attn	MNLI	.9440	.2391	.2523	.2549	.2473	.2370
	Quora	.9427	.0888	.0143	.0894	.0182	.1017
	SNLI	.9494	.3047	.2566	.3158	.2517	.2938
	IMDb	.8155	.1031	.2188	.2494	.2209	.2309
	SST-2	.7963	.1369	.1093	.1372	.1101	.1400
||Attn||	MNLI		.2396	.2576	.2586	.2518	.2409
	Quora		.0876	.0091	.0803	.0134	.0936
	SNLI		.3013	.2546	.3124	.2499	.2912
	IMDb		.0989	.2143	.2423	.2163	.2255
	SST-2		.1722	.1514	.1733	.1523	.1679
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
IntGrad	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
GradSHAP	MNLI						.4015
	Quora						.2433
	SNLI						.2219
	IMDb						.7021
	SST-2						.8056

Table 4: Mean Kendall-τ correlation between each pair of explanation methods for the **BiLSTM** model with **Softmax** attention applied to 500 randomly selected instances from the test portion of each dataset. Values are averaged across three independently seeded runs. **Key**: Attn = Attention Weights, ||Attn...|| denotes weighted vector norm analysis.

		Attn Roll	||Attn Roll||	||Attn Flow||	LIME	IntGrad	DeepLIFT	GradSHAP	DeepSHAP
Attn Flow	MNLI	.3344	.6593	.8214	.1121	.1264	.2118	.1183	.2051
	Quora	.3407	.6439	.7868	.1255	.1209	.2253	.0981	.2224
	SNLI	.3934	.7203	.7283	.1257	.1470	.2150	.1117	.1956
	IMDb								
	SST-2	.4256	.6591	.7663	.1393	.1001	.1214	.0903	.0961
Attn Roll	MNLI		.5920	.2919	.1206	.1718	.1812	.1760	.1693
	Quora		.5902	.3182	.0491	.1934	.0975	.1625	.1124
	SNLI		.5630	.2585	.0608	.2072	.1110	.1694	.0808
	IMDb		.6184		.0691	.1376	.2306	.1186	.1871
	SST-2		.6100	.3591	.0606	.1084	.1213	.1092	.1051
||Attn Roll||	MNLI			.6511	.1367	.1677	.2308	.1676	.2061
	Quora			.6481	.1153	.1877	.2053	.1533	.2049
	SNLI			.6323	.1264	.1797	.2164	.1436	.1746
	IMDb				.0704	.1293	.2411	.1099	.1988
	SST-2			.6672	.1069	.1076	.1495	.1032	.1202
||Attn Flow||	MNLI				.1034	.1141	.1966	.1064	.1870
	Quora				.1235	.1185	.2066	.0989	.1993
	SNLI				.0998	.0931	.1799	.0717	.1767
	IMDb								
	SST-2				.1391	.0854	.1295	.0787	.1027
Lime	MNLI					.0996	.1151	.0916	.0905
	Quora					.0884	.0956	.0651	.0854
	SNLI					.0818	.1248	.0573	.0906
	IMDb					.0525	.0544	.0469	.0473
	SST-2					.1422	.0815	.1058	.0600
IntGrad	MNLI						.2045	.4578	.1595
	Quora						.1046	.4615	.1023
	SNLI						.1146	.3737	.0757
	IMDb						.1429	.5020	.1180
	SST-2						.0539	.4363	.0509
DeepLIFT	MNLI							.2103	.4923
	Quora							.0920	.6512
	SNLI							.1094	.5089
	IMDb							.1292	.4332
	SST-2							.0600	.4361
GradSHAP	MNLI								.1626
	Quora								.0900
	SNLI								.0817
	IMDb								.1019
	SST-2								.0568

Table 5: Mean Kendall-τ correlation between each pair of explanation methods for the **DistilBERT** model with $\boldsymbol{\alpha}$**-entmax** self-attention applied to 500 randomly selected instances from the test portion of each dataset. Values are averaged across three independently seeded runs. **Key**: Attn Flow = Attention Flow, Attn Rollout = Attention Rollout, ||Attn...|| denotes weighted vector norm analysis.

		Attn Roll	||Attn Roll||	||Attn Flow||	LIME	IntGrad	DeepLIFT	GradSHAP	DeepSHAP
Attn Flow	MNLI	.4010	.6760	.8151	.1146	.1251	.2159	.1227	.2148
	Quora	.6318	.7612	.8095	.1117	.0367	.2426	.0241	.2319
	SNLI	.4549	.7229	.7737	.1030	.0753	.2178	.0571	.2149
	IMDb								
	SST-2	.6092	.6939	.7225	.1639	.0689	.1286	.0811	.1202
Attn Roll	MNLI		.6168	.3691	.1595	.1891	.2432	.1905	.2067
	Quora		.6849	.5587	.0992	.0574	.2267	.0518	.2257
	SNLI		.5567	.3604	.1048	.1645	.2214	.1600	.1796
	IMDb		.6895		.0991	.1818	.2516	.1432	.2303
	SST-2		.6249	.4791	.1271	.0511	.1328	.0737	.1291
||Attn Roll||	MNLI			.6969	.1427	.1671	.2332	.1665	.2150
	Quora			.7606	.1047	.0514	.2318	.0431	.2243
	SNLI			.7022	.1313	.1172	.2489	.1036	.2081
	IMDb				.0898	.1598	.2359	.1253	.2118
	SST-2			.7221	.1443	.0707	.1367	.0811	.1184
||Attn Flow||	MNLI				.1047	.1189	.1981	.1185	.1948
	Quora				.1039	.0447	.2239	.0332	.2127
	SNLI				.0899	.0587	.1940	.0424	.1908
	IMDb								
	SST-2				.1379	.0631	.1246	.0688	.1066
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
IntGrad	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
GradSHAP	MNLI								.1752
	Quora								.0535
	SNLI								.0851
	IMDb								.1093
	SST-2								.0419

Table 6: Mean Kendall-τ correlation between each pair of explanation methods for the **DistilBERT** model with **Softmax** self-attention applied to 500 randomly selected instances from the test portion of each dataset. Values are averaged across three independently seeded runs. **Key**: Attn Flow = Attention Flow, Attn Rollout = Attention Rollout, ||Attn...|| denotes weighted vector norm analysis.

B ACL Submission

B.1 Paper

Order in the Court: Explainable AI Methods Prone to Disagreement

Anonymous ACL-IJCNLP submission

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 appendix and 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. 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, `xai-agreement`, will be made publicly available upon publication.

	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.

4.5 Explaining the models

We leverage existing implementations of Integrated Gradients, DeepLIFT, Grad-SHAP, and Deep-SHAP[2] 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] https://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	.1849
	Quora	.0888	.0143	.0894	.0182	.1017	
	SNLI	.3047	.2566	.3158	.2517	.2938	
	IMDb	.1031	.2188	.2494	.2209	.2309	
	SST-2	.1369	.1093	.1372	.1101	.1400	
LIME	MNLI		.2535	.2176	.2488	.1923	**.3530**
	Quora		.1064	.1633	.1117	.1357	
	SNLI		.2232	.1790	.2156	.1646	
	IMDb		.2514	.2397	.2505	.2326	
	SST-2		.6230	.5921	.6228	.5538	
Int-Grad	MNLI			.4984	Grad-SHAP	.4015	
	Quora			.2906		.2433	
	SNLI			.2461		.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	**.1560**
	Quora	.0992	.0574	.2267	.0518	.2257	
	SNLI	.1048	.1645	.2214	.1600	.1796	
	IMDb	.0991	.1818	.2516	.1432	.2303	
	SST-2	.1271	.0511	.1328	.0737	.1291	
LIME	MNLI		.1037	.1228	.0969	.1078	.0928
	Quora		.0512	.0809	.0394	.0699	
	SNLI		.0598	.1351	.0550	.0969	
	IMDb		.0775	.0596	.0707	.0558	
	SST-2		.1869	.0726	.1571	.0534	
Int-Grad	MNLI			.2153	Grad-SHAP	.1752	
	Quora			.0625		.0535	
	SNLI			.0955		.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).

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 cross-lingual 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.

A Reproducibility Checklist

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

A.1 For all reported experimental results

A.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.

A.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)

We have submitted a zip file of our Python package, `xai-agreement`, which contains the source code to reproduce our experiments. Upon publication, we will publicly release this package on GitHub.

A.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`.

A.1.4 Average runtime for each approach

Refer to table 3 for the average time to train each model on each dataset.

A.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.

A.1.6 Corresponding validation performance for each reported test result

Table 4 details the validation performance of the best model weights for each dataset.

A.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 included in our submitted zip file.

B For all experiments with hyperparameter search

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

C For all datasets used

C.0.1 Relevant statistics such as number of examples

Table 5 lists the number of instances in each split of each dataset.

C.0.2 Details of train/validation/test splits

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

C.0.3 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.

C.0.4 A link to a downloadable version of the data

Links to download versions of all datasets are included in the README of the `xai-agreement` Python package. As discussed in section 4.1, we use the same splits as (Jain and Wallace, 2019) for the SST-2 and IMDb datasets. For posterity, links to all datasets are listed here: **SST-2**[3], **IMDb**[4], **SNLI**[5], **MNLI**[6], XNLI[7]. We have attached the **Quora** Question Pair dataset to this submission and will make it publicly available on release.

C.0.5 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.

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

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

[5] https://nlp.stanford.edu/projects/snli/

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

[7] 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 3: Number of minutes (average $\pm$ standard deviation) required to train each model on each dataset reported across three seeds.

	BiLSTM	DistilBERT
MNLI	67.088 ± 0.19	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 4: Validation accuracy (average $\pm$ standard deviation) of the selected model epoch reported across three seeds.

	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 5: 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.

C EACL Submission

C.1 Paper

Order in the Court: Explainable AI Methods Prone to Disagreement

Anonymous EACL submission

Abstract

Given a model and its corresponding decision, a subset of eXplainable AI (XAI) methods produce an *explanation* by quantifying each input's contribution toward that decision. The correlation between their generated explanations characterizes the extent to which XAI methods agree. In Natural Language Processing, one reason why *attention-based explanations* are discouraged is that they can disagree with some established XAI methods. For a more comprehensive comparison, we study the agreement of more recent XAI methods and attention-based explanations by comparing the Kendall-τ correlation of their full input token rankings across neural architectures and classification tasks. We find that none of our chosen XAI methods agree for a Transformer-based model. We argue that an XAI method's suitability is model and task-specific. Therefore, we urge practitioners to select appropriate XAI methods carefully.

1 Introduction

The notion of justice and, by extension, fairness is a fundamental virtue shared across ages and cultures (Peterson et al., 2004). We are more likely to accept a verdict as fair when the provided *explanation* is *faithful* to the arbiter's decision process and *plausible* to our own. Alarmingly, as deep neural models increasingly make medical, financial, and legal decisions, criteria for *plausible* and *faithful explanations* of their predictions remain speculative (Lipton, 2018; Jacovi and Goldberg, 2020).

Methods in eXplainable AI (XAI) offer *explanations* for model predictions, typically by quantifying each input's contribution toward that prediction. Additivity — treating all contributions as independent and quantifiable — is a common simplifying assumption. In this work, we focus on such additive methods and refer to them with the general term *XAI methods*. We denote the contribution of each input toward the model's decision as its *importance*. We say that XAI methods *agree* if there is a strong correlation between their computed rankings of input importance.

In Natural Language Processing, *attention-based explanations* assign scores from the weight distribution of the attention mechanism over an input sequence of tokens. Attention-based explanations, often visualized as heatmaps, are sometimes presented as an XAI method without justification (e.g., Martins and Astudillo, 2016; Wang et al., 2016). Jain and Wallace (2019) declare attention-based explanations unsuitable because they do not agree with several established XAI methods. Disqualifying a potential XAI method for disagreeing with other methods assumes the existence of an unobserved *true* explanation with which all valid XAI methods align. In this work, we hold a selection of more recent XAI methods to the same standard by measuring their agreement. Specifically, we answer 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 (recurrent and Transformer-based), or* ***(b)*** *the nature of the classification task (single- and pair-sequence)?*

We show that none of our chosen XAI methods agree when applied to our Transformer-based model and postulate that agreement declines as task and model complexity increases. We report higher overall agreement with attention-based explanations than with the XAI methods for our Transformer-based model. For these reasons, we recommend careful selection of appropriate XAI methods depending on the model and task.

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. They apply a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) model with an additive ($tanh$) (Bahdanau et al., 2015) attention mechanism to single and pair-sequence classification tasks. They justify one component of their conclusion that 'Attention is not Explanation' based on these findings. We build upon their work by comparing against more recent XAI methods and by including a Transformer-based model on the same types of tasks.

Recent efforts seek to enhance the faithfulness and plausibility of attention-based explanations. Typically, success is demonstrated by an increased agreement with XAI methods. One direction directly modifies the attention mechanism, either through guided training (Zhong et al., 2019), sparsity (Correia et al., 2019), or the minimization of hidden state conicity (Mohankumar et al., 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). Typically, agreement is measured with a small set of XAI methods. In contrast, we perform a more comprehensive comparison with 5 recent 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

In this section, we introduce our chosen XAI methods. We refer to a *candidate instance* as an instance that requires an explanation. Where applicable, we obtain token-level attribution by summing over the attribution to their individual features.

LIME (Ribeiro et al., 2016) produces locally faithful explanations by learning an easily explained (e.g., linear) model from samples weighted by their proximity to the candidate instance. **Integrated Gradients** (Sundararajan et al., 2017) calculates input feature attributions by accumulating the gradients obtained from the model along the straight-line path from a baseline to the candidate instance. **DeepLIFT** (Shrikumar et al., 2017) also produces input feature attributions using the gradients, but it assigns scores based on the difference between a reference activation and the activation of the candidate instance. This allows the calculated contributions to remain non-zero even when the gradients are zero. **SHAP** (Lundberg and Lee, 2017) identifies a unique solution for the contribution of each input toward the prediction. Since this is computationally expensive, they propose approximations based on existing methods: **Grad-SHAP** (Integrated Gradients), and **Deep-SHAP** (DeepLIFT).

3.2 Attention-based explanation

We define an *attention-based explanation* as an assignment of attention weights α over an input sequence of tokens. Since the dimensionality of α 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 that tokens' identities are linearly combined through the layers based exclusively on attention weights. Thus, the attention from any layer to the next is obtained via recursive matrix multiplication of weight matrices in the lower layers. Like Abnar and Zuidema (2020), we use the attribution calculated for the final layer's [CLS] token.

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 the work of Jain and Wallace (2019), we use identical splits and pre-processing techniques. 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 greater than 200 tokens. Readers may refer to our codebase for further details Most importantly, we contextualize the attention mechanism's utility for each dataset by comparing 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

For our experiments, we prefer a decrease in runtime over a minimal increase in accuracy on the downstream task. Therefore, we fine-tune the lighter, pre-trained DistilBERT model (Sanh et al., 2019) with 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. We use a patience value of 5 epochs for early stopping. For the BiLSTM, we follow the suggestions of 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. Our extendable framework for evaluating agreement between XAI methods and attention-based explanations, `xai-agreement`, will be made publicly available upon publication.

	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	$.823 \pm .008$	$.826 \pm .011$	$.823 \pm .004$	$.842 \pm .003$

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.

4.5 Explaining the models

We leverage existing implementations of Integrated Gradients, DeepLIFT, Grad-SHAP, and Deep-SHAP[1] 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 the XAI methods to 500 random instances taken from the test set of the corresponding dataset.

5 Results

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 that our XAI methods neither agree with each other (average=0.2229) nor with attention-based explanations (average=0.1705) across all models and tasks. **XAI methods do not correlate well with each other or with attention-based explanations**.

[1] https://github.com/pytorch/captum

For **RQ(a)**, the agreement between the recent XAI methods is much lower for the DistilBERT model (average=0.0928) than for the BiLSTM model (average=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). Thus, we see that **the amount of correlation depends on the model and task.**

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

(a) BiLSTM

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

The additivity assumption of XAI methods may be plausible in a single-sequence task like binary sentiment classification, where the removal of a negation term can flip the prediction. In pair-sequence tasks, however, input interactions may be complex and non-linear. As the complexity of interactions increases, it stands to reason that different additive approximations may capture distinct parts of the decision process. This could explain why the correlation decreases on pair-sequence tasks, in particular for the DistilBERT model, whose architecture allows all tokens to interact with each other.

For DistillBERT, we see higher overall agreement with attention rollout than with our XAI methods. One possible reason is that an attention-based explanation may capture an attention-based model's reasoning process more faithfully (Jacovi and Goldberg, 2020). This notion may also explain the low agreement for the BiLSTM's attention-based explanations: since the model's accuracy does not depend on the attention mechanism in most cases, it is likely not a meaningful part of its inference process. Our results suggest that faithfulness may be task-specific as well. We observe relatively high agreement between our model's attention-based explanations and the XAI methods on the SNLI and MultiNLI datasets. Notably, these datasets are also where the attention mechanism proves most useful.

In conclusion, we report a concerning lack of agreement between the XAI methods applied to our Transformer-based model, DistillBERT, indicating that perhaps the complexity of the explanation should more faithfully match the complexity of the decision process. If there is an underlying *true* explanation, then all of our XAI methods miss the mark. Furthermore, since attention rollout is the most agreeable explanation method for DistillBERT, we cannot say that 'Attention is not Explanation.' Instead, we recommend practitioners ensure the assumptions made by XAI methods are compatible with their model and task.

In future work, we would like to study the agreement between the top-k features selected by XAI methods since these may be of most interest to the end-user. Top-k token comparison to human-annotated rationales is common in NLP (DeYoung et al., 2020; Atanasova et al., 2020) and has demonstrated success when applied to attention-based explanations (Treviso and Martins, 2020).

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. *CoRR*, abs/2009.13295.

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*.

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.

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 cross-lingual 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.

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.

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.

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 computation*, 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.

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.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

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.

André F. T. Martins and Ramón F. Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1614–1623. JMLR.org.

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.

C. Peterson, M.E.P. Seligman, F.L.P.P.M.E.P. Seligman, American Psychological Association, and Oxford University Press. 2004. *Character Strengths and Virtues: A Handbook and Classification*. v. 1. American Psychological Association.

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 V. Treviso and André F. T. Martins. 2020. Towards prediction explainability through sparse communication. *CoRR*, abs/2004.13876.

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.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. 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.

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.

D Soos Talk Slides

All Roads Lead to Rome: Rank Correlation does not Measure the Utility of Feature Additive Methods in Explainable AI

Stefan Schouten & Michael Neely

Supervision: Maurits Bleeker & Ana Lucic

University of Amsterdam

February 26, 2021

Introduction - Why We're Here

To present our research that compares a number of explainability methods in a variety of settings.

- Is Attention a kind of Explanation?
- What makes an acceptable explanation?
- What assumptions underpin the comparison of feature-additive methods?

Introduction - Who We Are

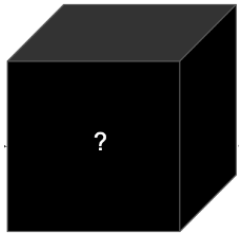
- Stefan Schouten
- Second Year MSc AI Student
- Thesis Area: Incorporating Semantics in Knowledge Graph Embeddings

- Michael Neely
- Second Year MSc AI Student
- Thesis Area: Multihop Question Answering

Outline

Explainable AI

- **Goal**: Make models transparent by explaining their predictions.
- **Desirable Properties** (Jacovi and Goldberg, 2020)
  - Faithful: to the model's reasoning process
  - Plausible: to human stakeholders



Post-Hoc Methods

- **Feature-based Explanations**
  - ⋆ *Feature-Additive*: Importance weights associated with input features (e.g., Ribeiro, Singh, and Guestrin, 2016; Lundberg and Lee, 2017; Sundararajan, Taly, and Yan, 2017).
  - *Minimal Sufficient Subsets*: Unique subset of input features that lead to the same decision (e.g., Chen et al., 2018; Yoon, Jordon, and Schaar, 2019).
- **Counterfactual Explanations**: highlight how a decision could be flipped (Wachter, Mittelstadt, and Russell, 2018).

What about Attention?

- Build a context vector to weigh the relevance of different input regions.
- Attention weights may be thought of as a feature-additive explanation method
- Advantages as an explanation method:
  - Faithful: Trained with the model and improves predictive power (Pruthi et al., 2020)
  - Plausible: Visualized with intuitive
  - Efficient: No extra computational cost

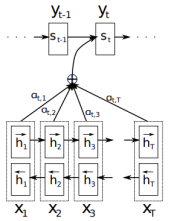
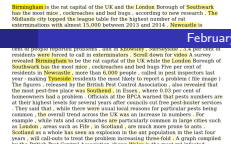


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Figure: Attention Mechanism from (Bahdanau, Cho, and Bengio, 2015).



The Origin of the Attention Debate

Jain and Wallace, 2019: If Attention is a valid form of explanation, it must satisfy two properties:

- **Property 1**: Different Attention weights should lead to different predictions.
- ⋆ **Property 2**: Attention weights should correlate (agree) with other feature-additive explainability methods.

Jain & Wallace Conclude: Neither property holds $\implies$ "Attention is not Explanation".

Example - Measuring Agreement

Task: Movie Review

Explanation Methods: LIME, Feature Ablation, Attention

	An	ungainly	,	humorless	rush	job	...
LIME	6th	4th	7th	1st	2nd	3rd	5th
Feature Ablation	6th	3rd	7th	2nd	1st	4th	5th
Attention	7th	1st	6th	2nd	5th	4th	3rd

τ(LIME, Feature Ablation) = .80 (strong correlation)

τ(LIME, Attention) = .33 (weak correlation)

τ(Feature Ablation, Attention) = .33 (weak correlation)

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

Agreement as an Evaluation Metric

- Has been used to invalidate Attention as an explanation method (Jain and Wallace, 2019)
- Has been used to justify alternative analyses of the Attention mechanism (Abnar and Zuidema, 2020; Kobayashi et al., 2020)

Is relative rank correlation (agreement) an appropriate evaluation metric for feature-additive explanation methods?

- We need to hold all feature-additive methods to the same standard

Research Question

How well do modern feature-additive methods (including Attention) agree with each other across multiple tasks and datasets? What can we conclude from these observations?

Methods

- Given input tokens $S = t_1, ..., t_n$ and a models 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

Experiments

- Two models:
  - Recurrent model: **BiLSTM** (Graves and Schmidhuber, 2005)
  - Transformer model: **DistilBERT** (Sanh et al., 2019)
- Two kinds of tasks:
  - **Single-sequence**: binary sentiment classification (SST, IMDb)
  - **Pair-sequence**: natural language inference (SNLI, MNLI) and paraphrase detection (Quora)
- Modern feature-additive methods:
  - **LIME** (Ribeiro, Singh, and Guestrin, 2016).
  - **Integrated Gradients** (Sundararajan, Taly, and Yan, 2017), **GradSHAP** (Lundberg and Lee, 2017)
  - **DeepLIFT** (Shrikumar, Greenside, and Kundaje, 2017), **DeepSHAP** (Lundberg and Lee, 2017)

Results

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

(a) BiLSTM

- High correlation for single-sequence tasks. (SST, IMDb)
- Low correlation for pair-sequence tasks. (MNLI, Quora, SNLI)
- Low correlation for Attention on all tasks.
- Low correlation everywhere.

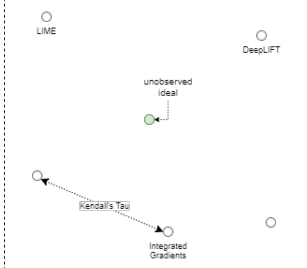
		LIME	Int-Grad	DeepLIFT	Grad-SHAP	Deep-SHAP	
Attn Roll	MNLI	.1595	.1891	.2432	.1905	.2067	.1560
	Quora	.0992	.0574	.2267	.0518	.2257	
	SNLI	.1048	.1645	.2214	.1600	.1796	
	IMDb	.0991	.1818	.2516	.1432	.2303	
	SST	.1271	.0511	.1328	.0737	.1291	
LIME	MNLI		.1037	.1228	.0969	.1078	.0928
	Quora		.0512	.0809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

(b) DistilBERT

Discussion I

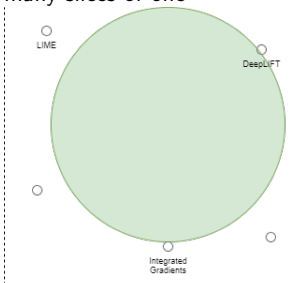
What does low agreement tell us?

- If we assume the existence of an ideal explanation, at most one method is "correct"



Low average agreement, single ideal explanation.

- Alternatively, there may be many "correct" explanations or at least many slices of one



Low average agreement, multiple ideal explanations.

Discussion I

What does high agreement tell us?

- It is unlikely the methods are harmonious in their error, since they use different algorithmic approaches
- However, we have **no external measure of quality**

High average agreement, single ideal explanation.

Discussion II

- Do rankings of feature importance capture the nuances of faithful and plausible explanations?
- Do these rankings make sense in more complex tasks (e.g., pair-sequence) where human annotators would struggle to assign per-token importance?

Main Takeaway

- Without making several (tenuous) assumptions, (low) agreement is often uninformative
- Investigating agreement on your model/task of interest is still valuable.
  If all methods agree then your choice between them becomes easy.

Future Work

- Agreement for top-k tokens (Treviso and Martins, 2020)
- Directly evaluate quality using human-annotated rationales (e.g., DeYoung et al., 2020; Atanasova et al., 2020)
- Pairwise interactions (Janizek, Sturmfels, and Lee, 2020)

Conclusion

- Thank you for your **attention** (heh).
- Questions?

References I

- Abnar, Samira and Willem Zuidema (July 2020). "Quantifying Attention Flow in Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4190–4197. DOI: 10.18653/v1/2020.acl-main.385. URL: https://www.aclweb.org/anthology/2020.acl-main.385.
- Atanasova, Pepa et al. (Nov. 2020). "A Diagnostic Study of Explainability Techniques for Text Classification". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 3256–3274. DOI: 10.18653/v1/2020.emnlp-main.263. URL: https://www.aclweb.org/anthology/2020.emnlp-main.263.
- Baan, Joris et al. (2019). "Understanding Multi-Head Attention in Abstractive Summarization". In: *CoRR* abs/1911.03898. arXiv: 1911.03898. URL: http://arxiv.org/abs/1911.03898.
- Bahdanau, Dzmitry, 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*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1409.0473.

References II

- Chen, Jianbo et al. (2018). *Learning to Explain: An Information-Theoretic Perspective on Model Interpretation*. arXiv: 1802.07814 [cs.LG].
- DeYoung, Jay et al. (July 2020). "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. URL: https://www.aclweb.org/anthology/2020.acl-main.408.
- Galassi, Andrea, Marco Lippi, and Paolo Torroni (2020). "Attention in Natural Language Processing". In: *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–18. ISSN: 2162-2388. DOI: 10.1109/tnnls.2020.3019893. URL: http://dx.doi.org/10.1109/TNNLS.2020.3019893.
- Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural Networks* 18.5. IJCNN 2005, pp. 602–610. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2005.06.042. URL: http://www.sciencedirect.com/science/article/pii/S0893608005001206.

References III

- Jacovi, Alon and Yoav Goldberg (July 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*. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://www.aclweb.org/anthology/2020.acl-main.386.
- Jain, Sarthak and Byron C. Wallace (June 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://www.aclweb.org/anthology/N19-1357.
- Janizek, Joseph D., Pascal Sturmfels, and Su-In Lee (2020). *Explaining Explanations: Axiomatic Feature Interactions for Deep Networks*. arXiv: 2002.04138 [cs.LG].

References IV

- Kobayashi, Goro et al. (Nov. 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)*. Online: Association for Computational Linguistics, pp. 7057–7075. DOI: 10.18653/v1/2020.emnlp-main.574. URL: https://www.aclweb.org/anthology/2020.emnlp-main.574.
- Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
- Pruthi, Danish et al. (July 2020). "Learning to Deceive with Attention-Based Explanations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4782–4793. DOI: 10.18653/v1/2020.acl-main.432. URL: https://www.aclweb.org/anthology/2020.acl-main.432.

References V

- Ribeiro, Marco Tulio, 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. San Francisco, California, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.
- Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108*.
- Shrikumar, Avanti, 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. Sydney, NSW, Australia: JMLR.org, pp. 3145–3153.
- Sundararajan, Mukund, 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. Sydney, NSW, Australia: JMLR.org, pp. 3319–3328.

References VI

- Treviso, Marcos and André F. T. Martins (Nov. 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*. Online: Association for Computational Linguistics, pp. 107–118. DOI: 10.18653/v1/2020.blackboxnlp-1.10. URL: https://www.aclweb.org/anthology/2020.blackboxnlp-1.10.
- Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2018). *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. arXiv: 1711.00399 [cs.AI].
- Yoon, Jinsung, James Jordon, and Mihaela van der Schaar (2019). "INVASE: Instance-wise Variable Selection using Neural Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BJg_roAcK7.

E FACT-AI Lecture Slides

Going Places With Your FACT-AI Work

Stefan Schouten & Michael Neely

University of Amsterdam

January 6, 2020

Introduction - Why We're Here

- To tell you about our experience in FACT AI 2019/2020
- To talk about how our experience lead us to take the Project AI elective and submit a paper to a NLP conference
- To talk about the content of that paper
- To encourage you to take full advantage of the opportunities FACT AI provides

Introduction - Who We Are

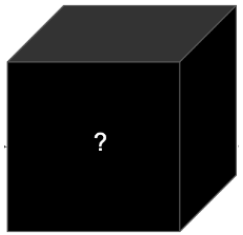
- Stefan Schouten
- Second Year MSc AI Student
- Thesis Area: Incorporating Semantics in Knowledge Graph Embeddings

- Michael Neely
- Second Year MSc AI Student
- Thesis Area: Multihop Question Answering

Outline

A Need for Transparency

- Despite recent focus on interpretability research, deep neural models remain (mostly) black boxes.
- Applying predictions from opaque models to real-world situations can be problematic, especially in medical, financial, and legal domains.
- The public may not be aware of limitations with these models.



Explainable AI

Goal: Make models transparent by explaining their predictions.
Post-Hoc Methods (non-exhaustive):

- **Feature-based Explanations**
  - *Feature-Additive*: Importance weights associated with input features (e.g., Ribeiro, Singh, and Guestrin, 2016; Lundberg and Lee, 2017; Sundararajan, Taly, and Yan, 2017).
  - *Minimal Sufficient Subsets*: Unique subset of input features that lead to the same decision (e.g., Chen et al., 2018; Yoon, Jordon, and Schaar, 2019).
- **Counterfactual Explanations**: highlight how a decision could be flipped (Wachter, Mittelstadt, and Russell, 2018).

Explainable AI

But what about an intrinsic method?

- **Faithful**: Trained *with* the model
  "a faithful interpretation is one that accurately represents the reasoning process behind the model's prediction." - Jacovi and Goldberg, 2020
- **Efficient**: No extra computational cost

What about attention?

Attention

- Build a context vector to weigh the relevance of different input regions (Bahdanau, Cho, and Bengio, 2015).
- Decoder can (soft-)search for the important sequence tokens (hopefully) necessary for prediction.
- Can thought of as a **feature-additive** explanation method for interpreting the behavior of (highly-complex) neural models.
- Used as a de facto explanation method for a few years (justified mostly with cherry-picked examples).

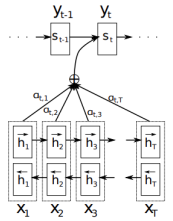
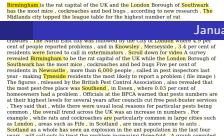


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Attention Mechanism from (Bahdanau, Cho, and Bengio, 2015).



Is Attention Explanation?

Attention is not Explanation (Jain and Wallace, 2019): our assigned paper in FACT AI 2019/2020.

Core argument: If Attention is a valid form of explanation, it must satisfy two properties:

- Property 1: Alternative Attention weight configurations should lead to different predictions
- Property 2: Attention weights should correlate with other additive feature importance methods

Conclusion: Neither property holds

Research Process - Literature Review

- Pruthi et al., 2020 agree with Property 1: "attention scores are easily manipulable"
- Wiegreffe and Pinter, 2019 dispute Property 1:
  - Existence does not entail exclusivity: adversarial weights decrease model performance
  - The attention distribution is not a primitive: cannot be manipulated post-hoc. Many other works fall into this trap (Serrano and Smith, 2019; Vashishth et al., 2019).
- We focused on Property 2 (Attention weights should correlate with other additive feature importance measures)

Attention is not Explanation

Experimental Specifics:

- Seq2seq encoder with additive Attention mechanism and feedforward decoder
- 3 encoder variants: **CNN** (Krizhevsky, Sutskever, and Hinton, 2012), **BiLSTM** (Graves and Schmidhuber, 2005), embedding average
- 2 attention types: **additive** (tanh), **scaled dot product**
- 3 tasks: binary text classification, question answering (QA), natural language inference (NLI)
- Additive feature importance measures: **input** $\times$ **gradient (Grad)** (Kindermans et al., 2016; Hechtlinger, 2016) and **leave-one-out (LOO)** (Li, Monroe, and Jurafsky, 2016)
- Correlation measure: **Kendall-τ** (Kendall, 1938)

Example - Sentiment Classification

- Example: "An ungainly, comedy-deficient, B-movie rush job..."
- Model: BiLSTM with additive (*tanh*) attention and whitespace tokenization
- Treat feature-additive scores as rankings
- Kendall-τ Correlation:
  - Attn vs LOO: $\sim$ 0.056 (no correlation)
  - Attn vs Grad: $\sim$ 0.056 (no correlation)
  - LOO vs Grad: $\sim$ 0.67 (strong correlation)
- Jain and Wallace, 2019 would say this is evidence against attention as an explanation method (Property 2)

token	Attn	LOO	Grad
an	.053	.033	.05
ungainly	.093	.222	.145
,	.10	.023	.046
comedy-deficient	.101	.001	.026
,	.10	.021	.048
b-movie	.129	.062	.034
rush	.156	.126	.069
job	.147	.122	.135
...	.123	0	.03

Research Process - Hypothesis Formulation

"We also acknowledge that irrelevant features may be contributing noise to the Kendall tau measure, thus depressing this metric artificially... it remains a possibility that agreement is strong between attention weights and feature importance scores for the top-k features only (the trouble would be defining this k and then measuring correlation between non-identical sets)" (Jain and Wallace, 2019).

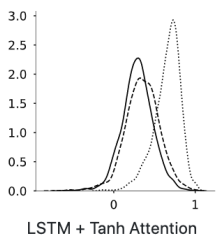
Research Process - Hypothesis Formulation

- Can we reproduce the paper's results when accounting for this hypothetical noise?
- How to choose the top-k input features?
- One possibility: **Sparsemax** (Martins and Astudillo, 2016)
  - Need to convert attention weights to a probability distribution
  - Traditional Softmax approach assigns non-zero mass to all hidden states
  - Sparsemax allows probabilities of zero
  - Model 'picks' k itself.
- Top-k features are those with non-zero probability assigned to their hidden representations
- Correlation measure: top-k kendall-τ (Fagin, Kumar, and Sivakumar, 2003)

Research Process - Experiments

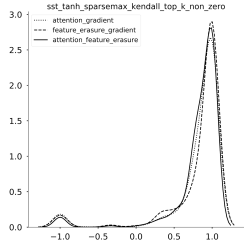
- Same BiLSTM model
- Binary Sentiment Classification task
- Datasets: Stanford Sentiment Treebank (Socher et al., 2013), IMDB (Maas et al., 2011)(same splits as J&W)
- Attention types: additive (*tanh*), scaled dot product
- Attention activation functions: Softmax and Sparsemax
- Feature Importance Measures: Leave-One-Out, Input $\times$ Gradient
- Correlation Measures
  - Kendall-τ
  - Top-k kendall-τ, k =smallest number of non-zero elements in either list
  - Top-k kendall-τ, k =average dataset sequence length

Research Process - Results



LSTM + Tanh Attention

Jain and Wallace, 2019: low correlation



Our result: High correlation

Research Process - Conclusion

- In certain settings, Attention may be a noisy form of explanation (evidence: high top-k correlation with feature importance measures)
- Rank correlation between feature-additive explanations depends on model, task, and dataset
- Faulty implementation of top-k Kendall-τ might be responsible for observations. (Not discovered until much later)

Project AI - Continuing our Research

- Future Work:
  - Other models and tasks
  - More contemporary feature-additive explanation methods
- Found two interested supervisors (PhD)
- Submitted our first project proposal (Project AI Page)
- **Research Question**: What are the trends in rank correlation between modern feature-additive methods (and attention) across multiple tasks and datasets?

Project AI - Fresh Perspective

- **Short paper**: focused narrative supported by solid evidence
- **Narrative**: Is relative rank correlation (agreement) an appropriate evaluation metric for feature-additive explanation methods?
  - Has been used to invalidate Attention as an explanation method (Jain and Wallace, 2019)
  - Has been used to justify alternative analyses of the Attention mechanism (Abnar and Zuidema, 2020; Kobayashi et al., 2020)
  - We need to hold other feature-additive methods to the same standard
- Submitted to NLP conference (acceptance decision pending)

Project AI - Experimental Setup

- Two models:
  - Recurrent model: **BiLSTM**
  - Transformer model: **DistilBERT** (Sanh et al., 2019)
    - Attention Rollout Abnar and Zuidema, 2020 to disentangle token attribution from the information mixing across layers in the multihead self-attention mechanism (Brunner et al., 2020)
- Two kinds of tasks:
  - **Single-sequence**: binary sentiment classification (SST, IMDb)
  - **Pair-sequence**: natural language inference (SNLI, MNLI) and paraphrase detection (Quora)
- Modern feature-additive methods:
  - Simplification-based: **LIME** (Ribeiro, Singh, and Guestrin, 2016).
  - Gradient-based: **Integrated Gradients** (Sundararajan, Taly, and Yan, 2017), **GradSHAP** (Lundberg and Lee, 2017)
  - Perturbation-based: **DeepLIFT** (Shrikumar, Greenside, and Kundaje, 2017), **DeepSHAP** (Lundberg and Lee, 2017)
- Correlation measure: **kendall-τ**

Project AI - Results

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

(a) BiLSTM

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

(b) DistilBERT

Project AI - Conclusion

- If the community insists on agreement to measure a feature-additive method's suitability, then all tested feature-additive methods fail for the Transformer, and, remarkably, attention rollout fails least strongly (it has the highest average agreement)
- Because feature-additive methods do not agree with each other, the choice between them is crucial. There is no ground truth for an explanation, and there isn't a particular feature-additive method that is demonstrably superior across all models and tasks.

Lessons Learned and Recommendations

- During FACT-AI, work on your project as if you will be submitting it for publication
- Find interested supervisors
- Submit project proposal 6-8 weeks before start date, maybe sooner
- If working in a team, make sure your thinking about unique contributions
- Develop a crisp narrative to guide your experiments (avoid scope creep)

Conclusion

- Thank you for your **attention** (heh).
- Questions?

References I

- Abnar, Samira and Willem Zuidema (July 2020). "Quantifying Attention Flow in Transformers". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4190–4197. DOI: 10.18653/v1/2020.acl-main.385. URL: https://www.aclweb.org/anthology/2020.acl-main.385.
- Baan, Joris et al. (2019). "Understanding Multi-Head Attention in Abstractive Summarization". In: *CoRR* abs/1911.03898. arXiv: 1911.03898. URL: http://arxiv.org/abs/1911.03898.
- Bahdanau, Dzmitry, 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*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1409.0473.
- Brunner, Gino et al. (2020). "On Identifiability in Transformers". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=BJg1f6EFDB.
- Chen, Jianbo et al. (2018). *Learning to Explain: An Information-Theoretic Perspective on Model Interpretation*. arXiv: 1802.07814 [cs.LG].

References II

- Fagin, Ronald, Ravi Kumar, and D. Sivakumar (2003). "Comparing Top k Lists". In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '03. Baltimore, Maryland: Society for Industrial and Applied Mathematics, pp. 28–36. ISBN: 0898715385.
- Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural Networks* 18.5. IJCNN 2005, pp. 602–610. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2005.06.042. URL: http://www.sciencedirect.com/science/article/pii/S0893608005001206.
- Hechtlinger, Yotam (2016). "Interpretation of Prediction Models Using the Input Gradient". In: *CoRR* abs/1611.07634. arXiv: 1611.07634. URL: http://arxiv.org/abs/1611.07634.
- Jacovi, Alon and Yoav Goldberg (July 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*. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://www.aclweb.org/anthology/2020.acl-main.386.

References III

- Jain, Sarthak and Byron C. Wallace (June 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://www.aclweb.org/anthology/N19-1357.
- Kendall, M. G. (1938). "A New Measure of Rank Correlation". In: *Biometrika* 30.1/2, pp. 81–93. ISSN: 00063444. URL: http://www.jstor.org/stable/2332226.
- Kindermans, Pieter-Jan et al. (2016). "Investigating the influence of noise and distractors on the interpretation of neural networks". In: *CoRR* abs/1611.07270. arXiv: 1611.07270. URL: http://arxiv.org/abs/1611.07270.
- Kobayashi, Goro et al. (Nov. 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)*. Online: Association for Computational Linguistics, pp. 7057–7075. DOI: 10.18653/v1/2020.emnlp-main.574. URL: https://www.aclweb.org/anthology/2020.emnlp-main.574.

References IV

- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., pp. 1097–1105. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
- Li, Jiwei, Will Monroe, and Dan Jurafsky (2016). "Understanding Neural Networks through Representation Erasure". In: *CoRR* abs/1612.08220. arXiv: 1612.08220. URL: http://arxiv.org/abs/1612.08220.
- Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
- Maas, Andrew L. et al. (June 2011). "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150. URL: https://www.aclweb.org/anthology/P11-1015.

References V

- Martins, André F. T. and Ramón Fernandez Astudillo (2016). *From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification*. arXiv: 1602.02068 [cs.CL].
- Pruthi, Danish et al. (July 2020). "Learning to Deceive with Attention-Based Explanations". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4782–4793. DOI: 10.18653/v1/2020.acl-main.432. URL: https://www.aclweb.org/anthology/2020.acl-main.432.
- Ribeiro, Marco Tulio, 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. San Francisco, California, USA: Association for Computing Machinery, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.
- Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108*.

References VI

- Serrano, Sofia and Noah A. Smith (July 2019). "Is Attention Interpretable?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2931–2951. DOI: 10.18653/v1/P19-1282. URL: https://www.aclweb.org/anthology/P19-1282.
- Shrikumar, Avanti, 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. Sydney, NSW, Australia: JMLR.org, pp. 3145–3153.
- Socher, Richard et al. (Aug. 2013). "Parsing with Compositional Vector Grammars". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 455–465. URL: https://www.aclweb.org/anthology/P13-1045.
- Sundararajan, Mukund, 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. Sydney, NSW, Australia: JMLR.org, pp. 3319–3328.
- Vashishth, Shikhar et al. (2019). *Attention Interpretability Across NLP Tasks*. arXiv: 1909.11218 [cs.CL].

References VII

- Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2018). *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. arXiv: 1711.00399 [cs.AI].
- Wiegreffe, Sarah and Yuval Pinter (Nov. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://www.aclweb.org/anthology/D19-1002.
- Yoon, Jinsung, James Jordon, and Mihaela van der Schaar (2019). "INVASE: Instance-wise Variable Selection using Neural Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BJg_roAcK7.