

---

# Reproducibility study - Does enforcing diversity in hidden states of LSTM-Attention models improve transparency?

---

Anonymous Author(s)

Affiliation

Address

email

## Reproducibility Summary

1

2 It has been shown (Jain and Wallace, 2019) that the weights in attention mechanisms do not necessarily offer a  
3 faithful explanation of the model's predictions. In the paper 'Towards Transparent and Explainable Attention Models'  
4 (Mohankumar et al., 2020) the authors propose two methods to enhance faithfulness and plausibility of the explanations  
5 provided by an LSTM model combined with a basic attention mechanism.

6 **Scope of Reproducibility** For this reproducibility study, we focus on the main claims made in this paper:

- 7
- 8 • The attention weights in standard LSTM attention models do not provide faithful and plausible explanations  
9 for its predictions. This is potentially because the concity of the LSTM hidden vectors is high.
  - 10 • Two methods can be applied to reduce concity: Orthogonalization and Diversity Driven Training. When  
11 applying these methods, the resulting attention weights offer more faithful and plausible explanations of the  
12 model's predictions, without sacrificing model performance.

12 **Methodology** The paper includes a link to a repository with the code used to generate its results. All our exper-  
13 iments with this code are conducted on GPU nodes of the Lisa Cluster at SURFsara<sup>1</sup>. We follow four investiga-  
14 tive routes: (i) Replication: we rerun experiments on datasets from the paper in order to replicate the results, and  
15 add the results that are missing in the paper; (ii) Code review: we scrutinize the code to validate its correctness;  
16 (iii) Evaluation methodology: we extend the set of evaluation metrics used in the paper with the LIME method, in an  
17 attempt to resolve inconclusive results; (iv) Generalization to other architectures: we test whether the authors' claims  
18 apply to variations of the base model (more complex forms of attention and a BiLSTM encoder)

19 **Results** We confirm that the Orthogonal and Diversity LSTM achieve similar accuracies as the Vanilla LSTM, while  
20 lowering concity. However, we cannot reproduce the results of several of the experiments in the paper that underlie  
21 their claim of better transparency. In addition, a close inspection of the code base reveals some potentially problematic  
22 inconsistencies. Despite this, under certain conditions, we do confirm that the Orthogonal and Diversity LSTM can  
23 be useful methods to increase transparency. How to formulate these conditions more generally remains unclear and  
24 deserves further research. The single input sequence tasks appear to benefit most from the methods. For these tasks, the  
25 attention mechanism does not play a critical role for achieving performance.

26 **What was easy / difficult** The codebase of the authors is accessible and can be run easily, with good facilities to  
27 prepare datasets and define configurations. The Orthogonalization and Diversity Training methods are well explained in  
28 the paper and mostly cleanly implemented. The larger datasets (Amazon and CNN) are difficult to run due to memory  
29 requirements and compute times. The codebase can be hard to navigate, a consequence of the choice to accommodate a  
30 large variation of models and datasets in one framework.

31 **Communication with original authors** We reached out to the authors on a fundamental but unexplained choice in  
32 the model architecture but unfortunately did not hear back before the deadline of our assignment.

---

<sup>1</sup>We had access to two Nvidia GTX1080Ti/11Gb VRAM GPUs. <https://userinfo.surfsara.nl/systems/lisa/description>.

## 33 1 Introduction

34 The popularity of attention models has sparked many studies on the interpretability of the attention distributions, with  
35 often conflicting claims (Jain and Wallace, 2019; Wiegrefe and Pinter, 2019; Serrano and Smith, 2019). Mohankumar  
36 et al. (2020) argue that the reason why attention weights do not always provide a faithful explanation of the model’s  
37 predictions is that the learned hidden states of the LSTM based encoder are very similar across time steps, which is  
38 expressed by high concity of these vectors. As a result, random permutation of the attention weights leads to a similar  
39 final context vector, which implies the weights do not provide a faithful explanation. The authors propose two methods  
40 that force the hidden states of the LSTM to be more diverse. Orthogonal LSTM ensures low concity by orthogonalizing  
41 the hidden state at time  $t$  with respect to the mean of the previous hidden states. In Diversity LSTM the model is trained  
42 to jointly maximize the log-likelihood of the training data and to minimize the concity of the hidden states.

## 43 2 Scope of reproducibility

44 In this reproducibility study we focus on the authors’ main claim that the Diversity LSTM and Orthogonal LSTM  
45 lead to more faithful and plausible explanations, while maintaining accuracy of the predictions. The authors support  
46 their claim by evaluating a series of metrics (Mohankumar et al., 2020) that are assumed to be indicative of levels of  
47 faithfulness and plausibility. We follow four investigative routes:

- 48 • Replication: The main part of our study is focused on reproducing the results on the metrics in Mohankumar  
49 et al. (2020), and to validate whether we can confirm their observations and conclusions. Furthermore, as the  
50 original paper only presents the results of a selection of models and datasets, we complement the results where  
51 possible. Most notably, we add results on the Orthogonal LSTM that were not in the original paper. Models,  
52 code and datasets are described in Section 3. Our replication results are presented in Section 4;
- 53 • Code review: As the authors’ code<sup>2</sup> is publicly available, we use their code for the reproduction. In Section 5  
54 we investigate whether the implementation is consistent with the description of the algorithms in the paper;
- 55 • Evaluation methodology: In Section 6 we report on our attempt to resolve inconclusive results we found on  
56 the attribution methods by extending the set of evaluation metrics used in the paper with the LIME method<sup>3</sup>;
- 57 • Generalization to other architectures: In Section 7 we test whether the authors’ claims apply to variations of  
58 the base model (more complex forms of attention and a BiLSTM encoder).

59 We conclude this paper in Section 8 with a discussion on the conditions under which the proposed methods are most  
60 likely to be effective, and a reflection on our replication study.

## 61 3 Methodology

62 **Code** The code accompanying the paper is an extension based on the code first developed by Jain and Wallace (2019)<sup>4</sup>.  
63 The entry point of the code is clear and well documented and allows a user to define specific jobs using command line  
64 arguments for hyperparameters. Preprocessing routines for the most datasets are included<sup>5</sup>.

65 **Datasets** We reran the experiments on 11 of the 14 datasets used in the paper. The nature and size of the datasets  
66 covers a wide range, from relatively simple binary sentiment classification tasks with single input sequence (abbreviated:  
67 SS) (e.g. SST with average input sentence length of 20 words), to complex question answering tasks with dual input  
68 sequences (abbreviated: DS)<sup>6</sup> (e.g. CNN with average document size of 760 words and an average of 26 answer  
69 categories). Some illustrations of data points can be found in Appendix D. The code repository includes links to the  
70 datasets, as well as the pre-processing routines used by the authors. We excluded the Amenia and Diabetes datasets  
71 because they were not accessible. The Amazon dataset caused memory issues when running the experiments. Despite  
72 these issues we were able to get the accuracies and concity values for this dataset.

73 **Model descriptions** The baseline model (Vanilla LSTM) used in the paper is shown in Figure 1. For DS tasks, it  
74 consists of two uni-directional LSTM encoders that act on a P-path (for document input phrases) and a Q-path (for

<sup>2</sup><https://github.com/akashkm99/Interpretable-Attention>

<sup>3</sup>The code with all extensions we made for this review can be accessed at [https://anonymous.4open.science/r/FACT\\_AI\\_project/](https://anonymous.4open.science/r/FACT_AI_project/)

<sup>4</sup><https://github.com/successar/AttentionExplanation>

<sup>5</sup>Diabetes and Anemia are not included. These datasets are not publicly available (ethics screening required before use)

<sup>6</sup>These distinctions are differently named in the code: SS is referred to as BC and DS as QA

75 question input phrases). When applied on SS tasks in the paper, only the P-path is used. An attention decoder is applied  
 76 to the hidden states of the P-path LSTM to form the context vector  $c_\alpha$  on which the model calculates its output. The  
 77 last hidden state of the Q-path is used as the query term for DS tasks.

78 The Diversity and Orthogonal LSTM that Mohankumar et al. (2020) propose are variants of the baseline model. The  
 79 **Orthogonal LSTM** applies an orthogonalization procedure to the LSTM hidden state vectors during training: the  
 80 hidden state in timestep  $t$  is set to the component that is orthogonal to the mean of previous hidden states. This  
 81 enforces low concity of the hidden state vectors  $h_t^p$ . The **Diversity LSTM** uses a standard LSTM cell with no explicit  
 82 orthogonalization, but minimizes concity jointly with the standard loss.

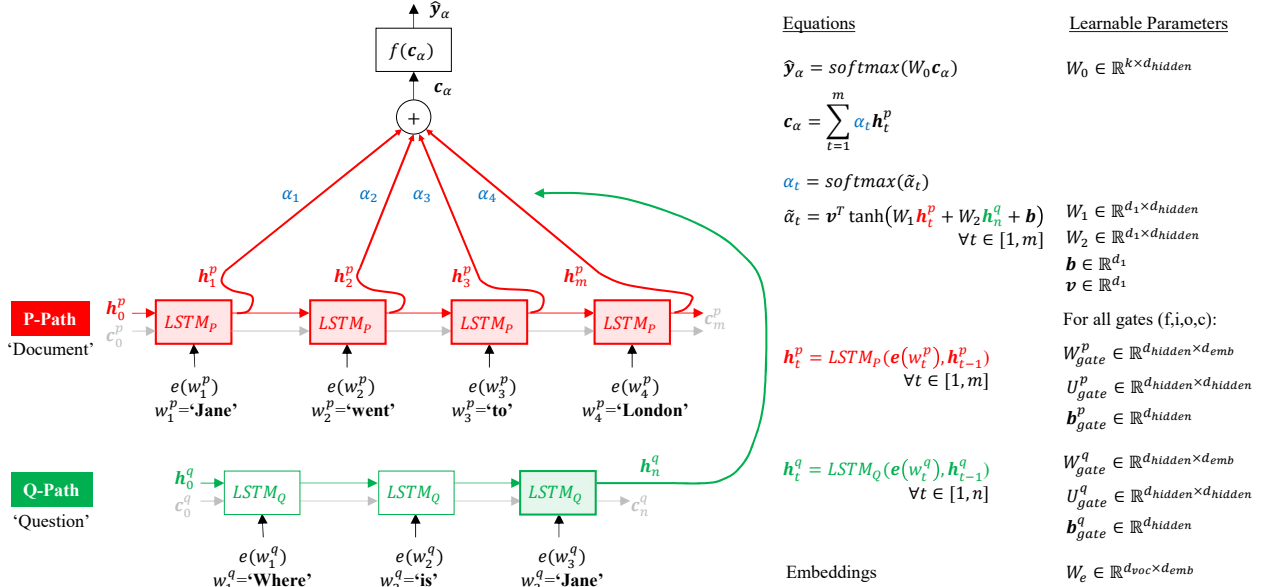


Figure 1: The LSTM+attention model as defined in the paper

83 **Hyperparameters** Given the wide variety of tasks and datasets, there is an elaborate set of model- and optimization  
 84 hyperparameters. Not all parameter values are indicated in the original paper, some were retrieved by inspecting the  
 85 code (an overview is presented in Appendix A). For all parameters, we used the defaults provided in the original code.  
 86 We do not engage in further hyperparameter optimization to stay close to the original paper’s approach. Note that we  
 87 are interested in transparency and explainability of the models, not their optimal performance.

88 **Experimental setup and computational requirements** We strictly follow the code environment as dictated by the  
 89 requirements file that accompanies the code. All models are run on Nvidia GTX1080Ti GPUs (11Gb VRAM).

90 Train and evaluation times varied between datasets and model variations, from ca. 5 minutes (SST dataset) to more than  
 91 40 hours (CNN dataset). We ran multiple seeds only on a selection of critical datasets to verify that differences we  
 92 observed w.r.t. the results in the original paper were significant. Due to resource constraints, all other comparisons  
 93 are based on single seeding, as was done in the original paper. This means that our observations are indicative, not  
 94 conclusive.

## 95 4 Replication of the paper’s results

### 96 4.1 Core replication results

97 Our reproduction study reveals numerous differences in results reported by Mohankumar et al. (2020), for all datasets  
 98 where we ran the experiments. Despite the differences, we support the observation that Diversity and Orthogonal  
 99 LSTM reach similar accuracies as Vanilla LSTM, and lower concity values, with the same exception reported in the  
 100 paper (CNN). However, we find the claim that Diversity LSTM leads to more transparent attention distributions is not  
 101 consistently supported. For Orthogonal LSTM, some results were omitted in the original paper, and we find conflicting  
 102 results about the effect on faithfulness and plausibility. We present an overview of the comparisons by metric, and the  
 103 impact our findings have on the main claims of the authors.

104 **Accuracy and concity** Of all accuracy and concity values reported by Mohankumar et al. (2020), we are able  
 105 to reproduce 86% within a 3%-point margin. Models and datasets that produced the most notable differences are  
 106 highlighted in Table 1. Despite the different values, the observation that Diversity and Orthogonal LSTM reach similar  
 107 accuracies as Vanilla LSTM still holds, except for the CNN dataset. Also, we can confirm that concity values are much  
 108 lower in Diversity and Orthogonal LSTM, except for CNN in the Diversity LSTM. The largest difference in accuracy  
 109 we observe for bAbI3, but the output files reveal that the model was not done training after the default 200 epochs.

Dataset	Accuracy%						Concity					
	Vanilla LSTM		Diversity LSTM		Orthogonal LSTM		Vanilla LSTM		Diversity LSTM		Orthogonal LSTM	
	original	rerun	original	rerun	original	rerun	original	rerun	original	rerun	original	rerun
SST	81.79	80.3	79.95	80.0	80.05	77.6	0.68	0.71	0.20	0.19	0.28	0.28
IMDB	89.49	89.3	88.54	87.8	88.71	88.3	<b>0.69</b>	<b>0.60</b>	0.08	0.09	0.18	0.16
Yelp	95.60	94.5	95.40	93.8	96.00	94.5	0.53	0.54	<b>0.06</b>	<b>0.35</b>	0.18	0.19
20News	93.55	90.8	91.03	90.8	92.15	91.9	0.77	0.76	0.15	0.14	0.23	0.24
Tweets	<b>87.02</b>	<b>83.3</b>	87.04	85.4	83.20	83.9	0.77	0.78	0.24	0.23	0.27	0.26
SNLI	78.23	77.3	76.96	74.0	76.46	76.6	0.56	0.59	<b>0.12</b>	<b>0.04</b>	<b>0.27</b>	<b>0.31</b>
QQP	78.74	78.4	78.40	78.2	78.61	78.6	0.59	0.58	0.04	0.03	0.33	0.32
bAbI1	99.10	100.0	100.00	100.0	99.90	99.9	<b>0.56</b>	<b>0.77</b>	0.07	0.07	0.22	0.23
bAbI2	<b>40.10</b>	<b>54.4</b>	<b>40.20</b>	<b>54.6</b>	56.10	59.0	<b>0.48</b>	<b>0.43</b>	<b>0.05</b>	<b>0.13</b>	<b>0.21</b>	<b>0.17</b>
bAbI3	<b>47.70</b>	<b>21.1</b>	<b>50.90</b>	<b>56.3</b>	<b>51.20</b>	<b>57.7</b>	<b>0.43</b>	<b>0.93</b>	0.10	0.11	0.12	0.13
CNN	<b>63.07</b>	<b>59.5</b>	<b>58.19</b>	<b>46.3</b>	54.30	53.6	<b>0.45</b>	<b>0.40</b>	<b>0.06</b>	<b>0.38</b>	0.07	0.10

Table 1: Comparison of reported accuracy and concity values (differences > 0.03 are highlighted).

110 **Importance of hidden representation** Mohankumar et al. (2020) analyse the importance of hidden representations  
 111 using intermediate representation erasure (Serrano and Smith, 2019) and also by examining the effect of permuting the  
 112 attentions weights (Jain and Wallace, 2019).

113 A visual comparison of the box plots about representation erasure in the paper with box plots in our reruns shows  
 114 similar results in 25 of the 30 boxes. Despite the fact that our rerun shows lower medians for the box plots for the  
 115 LSTM in IMDB and 20News dataset, the observation still holds that Diversity LSTM and Orthogonal LSTM reach a  
 116 quicker decision flip for SS tasks. We concur with the authors’ observations on the paraphrase detection (QQP) and  
 117 Q&A task (bAbI1). In our rerun we see that the quick decision flip that is shown in bAbI1 also occurs in bAbI2 and  
 118 bAbI3. Mohankumar et al. (2020) do not report on SNLI and CNN, where our rerun shows no improvement of the  
 119 Diversity LSTM and Orthogonal LSTM models over Vanilla LSTM.

120 The impact of permuting attention weights is difficult to compare with our results as Mohankumar et al. (2020) only  
 121 report a graphical representation (violin plots) of median output difference. After visual comparison we judge that the  
 122 overall results are similar for IMDB, 20News and Yelp. We also evaluate the median output difference for datasets not  
 123 reported by Mohankumar et al. (2020). We observe that the results for SST and Tweets show a similar ‘shift to the right’  
 124 as reported for other binary classification tasks. For DS tasks we observe that Vanilla LSTM already has relatively high  
 125 median output difference, and the Diversity LSTM and Orthogonal LSTM provide less improvement.

126 We conclude that in our experiments, the Diversity and Orthogonal LSTM do result in quicker decision flips and higher  
 127 output difference for SS tasks, but not consistently for the other tasks.

128 **Comparison with rationales** Our rerun of rationale length and rationale attention shows very different results as  
 129 reported by Mohankumar et al. (2020), see Table 2. Although we can confirm that Diversity LSTM results in shorter  
 130 rationales, we cannot support the claim that Diversity LSTM provides much higher attention to the rationale than  
 131 Vanilla LSTM. In our rerun this only holds for 20News.

132 The data for the Orthogonal LSTM, which were not reported by Mohankumar et al. (2020), show much shorter rationale  
 133 length, consistent with the paper’s claim. However, impact on the share of attention on the rationale is mixed: it is  
 134 higher for Yelp and 20 News, similar for IMDB and Tweets, but lower for SST.

135 For DS tasks, the rationale comparison is not implemented by the authors, we suspect because of the high computational  
 136 costs involved for calculating rationales in tasks with multiple output categories.

137 **Comparison with attribution methods** The rerun of the correlation metrics shows numerous differences in both  
 138 Pearson correlation and JS Divergence. After studying Pearson correlation, we support the authors’ claim that compared  
 139 with Vanilla LSTM, Diversity LSTM produces attention weights that better correlate with gradients and integrated  
 140 gradients, although in our results the relative increase of correlation with gradients is smaller: 13% instead of the 65%  
 141 reported by Mohankumar et al. (2020)<sup>7</sup>. However, we do not see the claimed reduction in JS Divergence. In fact, for

<sup>7</sup>This percentage represents the average of the increases over all datasets.

Dataset	Rationale attention						Rationale length					
	Vanilla LSTM		Diversity LSTM		Orthogonal LSTM		Vanilla LSTM		Diversity LSTM		Orthogonal LSTM	
	original	rerun	original	rerun	original	rerun	original	rerun	original	rerun	original	rerun
SST	<b>0.348</b>	<b>0.74</b>	0.624	0.55	-	0.35	<b>0.240</b>	<b>0.72</b>	0.175	0.18	-	0.10
IMBD	<b>0.472</b>	<b>0.97</b>	<b>0.761</b>	<b>0.91</b>	-	0.92	<b>0.217</b>	<b>0.92</b>	<b>0.169</b>	<b>0.22</b>	-	0.27
Yelp	0.438	0.43	<b>0.574</b>	<b>0.27</b>	-	0.55	<b>0.173</b>	<b>0.38</b>	0.160	0.19	-	0.11
20News	0.627	0.62	0.884	0.94	-	0.86	<b>0.215</b>	<b>0.59</b>	<b>0.173</b>	<b>0.27</b>	-	0.24
Tweets	<b>0.284</b>	<b>0.82</b>	<b>0.764</b>	<b>0.59</b>	-	0.79	<b>0.225</b>	<b>0.81</b>	0.306	0.32	-	0.39

Table 2: Comparison of reported rationales (differences > 0.05 are highlighted)

142 all datasets the Diversity LSTM produces similar or even higher JS Divergence values than Vanilla LSTM, except JS  
143 Divergence with Integrated Gradients for 20News, see Table 3. The Orthogonal LSTM, for which no correlation data is  
144 reported in the paper, is in line with the Diversity LSTM in this respect.

Dataset	JS Divergence Gradients						JS Divergence Integrated Gradients					
	Vanilla LSTM		Diversity LSTM		Orthogonal LSTM		Vanilla LSTM		Diversity LSTM		Orthogonal LSTM	
	original	rerun	original	rerun	original	rerun	original	rerun	original	rerun	original	rerun
SST	0.10	0.09	0.08	0.09	-	0.14	0.12	0.10	0.09	0.10	-	0.15
IMDB	0.09	0.08	0.09	0.11	-	0.13	0.13	0.11	0.13	0.15	-	0.18
Yelp	0.15	0.12	<b>0.13</b>	<b>0.17</b>	-	0.16	0.19	0.18	0.19	0.19	-	0.17
20News	0.15	0.18	<b>0.06</b>	<b>0.17</b>	-	0.17	0.21	0.22	<b>0.07</b>	<b>0.13</b>	-	0.15
Tweets	0.08	0.07	0.12	0.09	-	0.18	0.08	0.08	<b>0.15</b>	<b>0.10</b>	-	0.19
SNLI	0.11	0.11	0.10	0.11	-	0.12	0.16	0.14	0.13	0.14	-	0.15
QQP	<b>0.15</b>	<b>0.10</b>	0.10	0.11	-	0.12	<b>0.19</b>	<b>0.15</b>	0.15	0.14	-	0.14
bAbI1	<b>0.33</b>	<b>0.12</b>	0.21	0.23	-	0.21	<b>0.43</b>	<b>0.25</b>	0.24	0.22	-	0.28
bAbI2	<b>0.53</b>	<b>0.39</b>	<b>0.23</b>	<b>0.40</b>	-	0.38	<b>0.58</b>	<b>0.51</b>	<b>0.19</b>	<b>0.58</b>	-	0.54
bAbI3	<b>0.46</b>	<b>0.26</b>	0.37	0.36	-	0.43	<b>0.64</b>	<b>0.35</b>	<b>0.41</b>	<b>0.64</b>	-	0.64
CNN	<b>0.22</b>	<b>0.16</b>	<b>0.17</b>	<b>0.34</b>	-	0.39	<b>0.30</b>	<b>0.23</b>	<b>0.21</b>	<b>0.51</b>	-	0.44

Table 3: Comparison of correlation metrics. Differences > 0.03 are highlighted

145 **Analysis by POS tags** A comparison of the importance that is attributed to various POS tags shows similar importance  
146 and ranking for the SST, 20News and Tweets datasets. For Yelp and QQP we get different outcomes. Most notably,  
147 with vanilla LSTM model for Yelp we see no attention given to punctuations (PUNC), for which Mohankumar et al.  
148 (2020) reports highest attention. For QQP, Mohankumar et al. (2020) reports 23% on PUNC, while we find only 9%.  
149 Our results indicate the improvements shown in POS tags are less clear than reported by Mohankumar et al. (2020).

150 **Human evaluation** We could not reproduce the human evaluation within the four-week time frame of our research.  
151 Mohankumar et al. (2020) reports convincing results, and we also believe human interpretation should play a key role  
152 in judging whether their methods improve transparency. We include some examples in Appendix D for this purpose.

## 153 4.2 Conclusion regarding reproducibility

154 Our findings are summarized in Table 4. We conclude that it is not immediately clear that Diversity LSTM and  
155 Orthogonal LSTM provide better transparency for all the studied datasets.

- 156 • The Orthogonal LSTM clearly leads to lower concity than Vanilla LSTM, but Mohankumar et al. (2020) show  
157 little evidence with other metrics that indicate higher faithfulness: of the 14 datasets, only 6 boxplots and  
158 4 violin charts are included. The results observed in our rerun are mixed. For example, Orthogonal LSTM  
159 works well for 20News, but for SNLI there is hardly any effect on the box plot, and also correlation/JSD with  
160 (integrated) gradients is worse.
- 161 • For Diversity LSTM, Mohankumar et al. (2020) show convincing evidence with substantial data. We observe  
162 similar trends in concity, and the impact of diversity training is clear in the box plots and violin charts for the  
163 binary classification tasks. However, for the tasks that require two input sequences like SNLI, bAbI2, CNN our  
164 rerun shows that Diversity LSTM does not contribute much to faithfulness and can lead to lower correlation  
165 with (integrated) gradients and higher JS Divergence.

## 166 5 Code Review

167 As part of the reproduction study, we familiarized ourselves with the code to understand how the model and the  
168 experiments had been implemented. We also scrutinized the code to check whether we could find a cause for the  
169 differences we found in the reported metrics. The code’s class architecture can accommodate a wide range of tasks,

Metric	Claim (with reference to paragraph number in Mohankumar et al. (2020))	Supported after rerun	Notes
Accuracy and concity	Diversity LSTM and Orthogonal LSTM achieve similar accuracies as Vanilla LSTM, but much lower concity (§5.2)	Yes	Except CNN in Diversity LSTM
Fraction of hidden representation required for decision flip (box plots)	Diversity LSTM and Orthogonal LSTM reach quicker decision flip (§5.3)	Yes (for models in paper)	Especially for BC tasks, somewhat for QQP; not for SNLI and QA tasks
Median Output difference on randomly permuting attention weights (violin charts)	Diversity LSTM and Orthogonal LSTM are more sensitive to random permutation of weights than Vanilla LSTM (§5.3)	Yes (for models in paper)	Clear difference for BC tasks, mixed picture for the dual-sequence tasks.
Rationale attention	Diversity LSTM provides much higher attention to rationales than Vanilla LSTM across the 8 Text classification datasets (§5.4)	No	Only true for 20News; No results reported on QA tasks
Rationale length	Diversity LSTM often provides shorter rationales than Vanilla LSTM (§5.4)	Yes	No results reported on QA tasks
Pearson correlation and JS divergence between distribution of attention and (integrated) gradients	Attention weights in Diversity LSTM better agree with gradients and integrated gradients than Vanilla LSTM (§5.5)	Mixed	Diversity LSTM has higher Pearson correlation, but similar or higher JS Divergence
Attention given to POS tags	Attention given to punctuation marks is significantly reduced on the Yelp, Amazon and QQP datasets (§5.6)	No	Not for Yelp, less clear for QQP
	Diversity LSTM gives much more attention to adjectives than Vanilla LSTM in the four sentiment analysis tasks (SST, IMDB, Yelp, Amazon) (§5.6)	Yes	True for SST and IMDB, but not for Yelp
Human evaluation of plausibility	Human evaluators prefer attention distribution of Diversity LSTM over Vanilla LSTM for Yelp, SNLI, QQP and bAbI1 (§5.7)	Not reproduced	Evaluation by only 15 people

Table 4: Evidence for authors’ claims after rerun

170 datasets and model configurations. While convenient, this also makes the codebase complex and susceptible to errors.  
 171 The code review revealed several debatable choices, of which the main ones are described below.

172 **Orthogonalization of Q-path in dual input sequence tasks** For DS tasks, we expect the orthogonalization procedure  
 173 to only be activated in the P-path (the path of the input document) of the model, as this is the path on which the attention  
 174 mechanism applies its weights  $\alpha_t$ . However, in the code, orthogonalization is *also* applied to the Q-path (the path of the  
 175 question phrase in the Q&A tasks, or the second input phrase in SNLI and QQP).

176 In our view, this introduces a potentially problematic effect. The attention mechanism uses only the last hidden state  
 177 vector  $\mathbf{h}_t^q$  as the query term. This representation for the last word in the sequence will only retain the vector component  
 178 orthogonal to the mean of the previous word representations, as a result of orthogonalization. We argue that the direction  
 179 of  $\mathbf{h}_t^q$  in the hidden space will represent the exclusive ‘change of meaning’ that the last word adds to the sequence. This  
 180 is not a problem in the bAbI tasks, where the prompt word in the question phrase is always the last word (e.g., ‘Where  
 181 is Jane’). But for longer questions where the prompt words appear earlier in the question, this may impede the attention  
 182 mechanism from finding the right prompt words.

183 In order to test this sensitivity, we conduct an experiment  
 184 for the simpler SS tasks. We apply orthogonalization during  
 185 training and compare model performance when i) attention  
 186 weights are left unconstrained vs. ii) all attention weights are  
 187 set to zero, except for the last hidden state. The result is shown  
 188 in Table 5. What is striking is the performance remains on  
 189 par (marked in green) for Vanilla LSTM when only attending  
 190 to the last hidden state, indicating the model performs well  
 191 without the attention mechanism. However, we observe a  
 192 performance drop of 10%-34% (absolute) when attention is  
 193 constrained for the Orthogonal LSTM (marked in red). Indeed,  
 194 it appears part of the information required for inference is lost.

195 How this effect impacts the results requires further study. It  
 196 may explain the accuracy drop from 63% (Vanilla LSTM) to  
 197 58%/54%(Diversity/Orthogonal LSTM) for CNN as reported  
 198 in Table 2 by Mohankumar et al. (2020). We have contacted  
 199 the authors to verify their intentions, but did not receive a response prior to submission of this reproduction study.

Dataset	Vanilla LSTM		Orthogonal LSTM	
	Base attention	last_only attention	Base attention	last_only attention
SST: accuracy (conicity)	0.803 (0.713)	<b>0.810</b> (0.763)	0.776 (0.283)	<b>0.583</b> (0.265)
IMDB: accuracy (conicity)	0.893 (0.602)	<b>0.876</b> (0.885)	0.883 (0.163)	<b>0.784</b> (.141)
20News: accuracy (conicity)	0.908 (0.761)	<b>0.857</b> (0.831)	0.919 (0.235)	<b>0.583</b> (0.395)
Tweets: accuracy (conicity)	0.833 (0.776)	<b>0.782</b> (0.798)	0.839 (0.260)	<b>0.712</b> (0.330)

Table 5: Demonstration of adverse effect of orthogonalization on the information content of the last hidden state vector (results reflect our experiments, not the original paper)

200 **Disparate calculation of final prediction** For DS tasks, in the code the final prediction layer is implemented as  $\hat{y} =$   
 201  $\text{softmax}(\mathbf{W}_r(\tanh(\mathbf{W}_p\mathbf{c}_\alpha + \mathbf{b}_p + \mathbf{W}_q\mathbf{h}_n^q + \mathbf{b}_q)) + \mathbf{b}_r)$ . This deviates from the prediction  $\hat{y} = \text{softmax}(\mathbf{W}_0\mathbf{c}_\alpha)$

described in Section 2.1 by Mohankumar et al. (2020)<sup>8</sup>. However, this does not affect the core architecture, namely LSTM and attention, so we did not modify the code or conduct further experiments.

**Fine-tuning of embeddings** The models use pre-trained embeddings except for the bAbI datasets. Words outside of the pre-trained embeddings’ vocabulary are initialized with zero-vectors. All embeddings are fine-tuned (i.e. trainable), independently for the P- and Q-paths for DS tasks. This is not mentioned in the original paper and this choice is questionable as it leads to an excessive number of trainable parameters (e.g., >40M for the CNN dataset, see Appendix A) and training time, while it is unlikely to be critical for the tasks.

**Definition of dev set for bAbI datasets** While pre-processing bAbI datasets, 15% of the train set is randomly selected to be used as dev set, resulting in much higher similarity between these two splits compared to the test set. As a result, the trained model is overfit on the train set, and we observe a large gap between dev and test accuracy.

## 6 Extension of the evaluation methods

As discussed in Section 4.1, our rerun of Pearson’s correlation and JS Divergence between attention weights and gradients/integrated gradients points towards a less convincing conclusion. We therefore also used the LIME framework (Ribeiro et al., 2016) as a third metric for comparing how transparent the attention weights are as explanations, as well as how much improvements are brought about by the Diversity and Orthogonal LSTM.

We use LIME to generate a score for the predicted class on each word-position in the sentence, which can then be compared with the attention weights. For calculating JS divergence we also rescaled the lime score so that the scores range from 0 to 1, and sums to 1 per sentence (i.e. similar to attention scores). The results are shown in Table 6, where we experimented with only a representative selection of datasets due to time and resource constraints.

Similar to our comparison of attention weights with gradient-based methods, Table 6 indicates Diversity and Orthogonal LSTM fail to produce explanations consistent with LIME. It is also not clear which statistical measure is best for comparing whether two explanation methods agree with each other. In several instances (e.g. 20News and Tweets), we observe an increase in Pearson’s correlation and an increase in JS Divergence at the same time when going from Vanilla LSTM to Orthogonal/Diversity LSTM models.

Dataset	Pearson’s Correlation			JS Divergence		
	Vanilla	Ortho.	Div.	Vanilla	Ortho.	Div.
IMDB	0.42	<b>0.33</b>	0.42	0.26	<b>0.44</b>	<b>0.42</b>
20News	0.30	<b>0.70</b>	<b>0.71</b>	0.22	<b>0.42</b>	<b>0.45</b>
Tweets	0.13	<b>0.38</b>	<b>0.43</b>	0.07	<b>0.33</b>	<b>0.18</b>
SNLI	0.24	0.22	0.23	0.15	0.15	0.12
bAbI1	0.69	0.67	<b>0.58</b>	0.42	0.38	0.46

Numbers that agree with expectations (higher correlation, lower JS Divergence) are highlighted in green, numbers opposite to expectations are highlighted in red.

Table 6: Correlation and JS Divergence between attention weights and LIME scores

## 7 Generalization to other model architectures

Despite the differences we found between our observations and the observations reported by (Mohankumar et al., 2020), we still see the potential value of the methods they propose. This is because we did observe sparser attention weights when using Diversity and Orthogonal LSTM, and because of the strong preference expressed for the Diversity LSTM in the human evaluations conducted by Mohankumar et al. (2020).

We therefore investigate how well these methods work in alternative settings. So far the Orthogonalization and Diversity Training methods are only tested on one-layer uni-directional LSTM models with attention. However, in many recent studies, BiLSTM-based attention models or Transformer models are used (Zhou and Wu, 2018; Lee et al., 2019; Aziz Sharfuddin et al., 2018). Similarly, more complex attention mechanisms such as self-attention and multi-head attention (Vaswani et al., 2017) gained popularity due to their superior performance. For this reason, we investigate whether the proposed methods can be extended to more complex models and whether the authors’ two main claims still apply.

**Extending to other attention mechanisms** The application of more advanced attention mechanisms (such as multi-head attention) poses a challenge because they produce more than one attention weight per word. It is not straightforward to generate explanations and word importance based on these weights. As a consequence, several of the evaluation metrics used by the authors cannot be applied in their current form. This would require making non-trivial design choices on how to combine multiple distributions of the attention weights. Further research is required to investigate this and whether existing methods such as Attention Flow and Attention Rollout (Abnar and Zuidema, 2020) can provide a resolution.

<sup>8</sup><https://github.com/akashkm99/Interpretable-Attention/blob/master/model/modules/Decoder.py#L101-L107>

250 **Extending to other architectures: BiLSTM Experiments** We replace the uni-directional LSTM in the model  
 251 (Figure 1) with a bi-directional LSTM. We choose the BiLSTM architecture, and not a Transformer based architecture,  
 252 as the latter requires dealing with the more advanced attention mechanisms discussed above.

253 In order to maintain the decoder’s complexity (the attention mechanism), we preserve the output dimension of the  
 254 LSTM. This requires halving the dimension of the hidden states, which also ensures that the number of trainable  
 255 weights of the BiLSTM is comparable to that of the unidirectional LSTM. For the Diversity BiLSTM, the same diversity  
 256 weights are used as in Mohankumar et al. (2020). The conicity term present in the loss function of the Diversity  
 257 BiLSTM is calculated based on the concatenated forward and backward hidden representations. Orthogonalization for  
 258 the Ortho BiLSTM is applied before concatenation of the forward and backward hidden states.

259  
 260 Results show that the application of the two methods proposed by Mohankumar et al. (2020) do not result in  
 261 performance loss and do lower conicity. However, on other metrics and across datasets, the picture is mixed like we saw  
 262 in our reproducibility results for the unidirectional LSTM, indicating the methods do not unconditionally improve  
 263 explanations. We will not discuss these results in detail, but conclude that it is indeed possible to extend the proposed  
 264 methods to BiLSTM attention models. Full results are included in Appendix C for completeness.

## 265 8 Discussion

266 Our reproduction shows that enforcing low conicity between the hidden states of an LSTM encoder does not guarantee  
 267 improved transparency in the studied datasets, at least not on the metrics used by Mohankumar et al. (2020). We find  
 268 the authors’ claim about improved transparency not generally applicable and under certain conditions their methods  
 269 even hurt accuracy. Still, the Diversity LSTM and Orthogonal LSTM do lead to improved metrics on some datasets,  
 270 and the human evaluation Mohankumar et al. (2020) conducted shows strong preference for the Diversity LSTM over  
 271 Vanilla LSTM. This raises the question under what conditions these methods should be applied.

272 **Conditions underlying effectiveness** One pattern that seems to  
 273 emerge is that the benefits of orthogonalizing or diversity training are  
 274 most apparent for the relatively simpler SS tasks. The potential to im-  
 275 prove faithfulness of the weights might be high in those cases as it not  
 276 a given that attention weights carry any meaning for these task.

277 For some tasks, the LSTM does not strictly need the attention mechanism  
 278 to perform well, as is shown in Table 7 when the attention mechanism  
 279 is constrained to be either uniform or attending to the last word only.  
 280 In contrast, the more difficult DS tasks do require the attention mecha-  
 281 nism in order to reach higher accuracies. This pattern is similar to that  
 282 described by Wiegrefe and Pinter (2019).

283 We suspect that there is a relation between a) how crucial the attention  
 284 mechanism is for performance in a given task, b) how much improvement  
 285 Orthogonal/Diversity LSTM can offer w.r.t. plausibility of the attention  
 286 weights for explaining the model’s outputs. This relationship, and the  
 287 conditions under which orthogonalization and diversity training offer the  
 288 best results, deserves additional investigation.

289 **Reflection on our replication study** A key insight we have gained is that even with access to the original code, exact  
 290 reproduction of the results is not guaranteed. We have not been able to find the cause of several differences in results.  
 291 The available time and hardware limited our possibilities to repeat these experiments with multiple seeds to find an  
 292 estimate of the variance of outcomes.

293 Another insight we gained is that the metrics concerning faithfulness and plausibility can be hard to interpret, as it is  
 294 deeply entangled with the nature of the dataset as well as the model implementation. To enable scalable development of  
 295 transparent AI models, reliable quantitative metrics are needed that can accurately approximate real humans’ judgement.  
 296 We believe further development of transparency metrics is an important area for further research to help build more  
 297 transparent models.

## 298 Acknowledgement

299 We’d like to thank [name to be added] for his guidance and insightful discussions.

Dataset	Base attention		Constrained attention	
	Reported	Repr.	uniform	last_only
SST	.818	.803	.800	.810
IMDB	.895	.893	.883	.876
Yelp	.956	.949	.950	.949
20News	.936	.908	.898	.857
Tweets	.870	.833	.833	.782
SNLI	.782	.773	.755	.759
QQP	.787	.784	.789	.792
bAb1	.991	1.00	.485	.729
bAb2	.401	.544	.315	.441
bAb3	.477	.211*	-	-
CNN	.631	.595	.424	.367

\* Reproduction failed, comparisons not applicable

Table 7: Impact on performance of the Vanilla LSTM when forcing uniform, first- and last only attention



## References

- 300  
301 Samira Abnar and Willem Zuidema. 2020. Quantifying Attention Flow in Transformers. In *Proceedings of the*  
302 *58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for  
303 Computational Linguistics.
- 304 A. Aziz Sharfuddin, M. Nafis Tihami, and M. Saiful Islam. 2018. A deep recurrent neural network with bilstm model  
305 for sentiment classification. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*,  
306 pages 1–4.
- 307 Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *CoRR*, abs/1902.10186.
- 308 Lung-Hao Lee, Yi Lu, Po-Han Chen, Po-Lei Lee, and Kuo-Kai Shyu. 2019. Ncuae at mediqa 2019: medical text  
309 inference using ensemble bert-bilstm-attention model. In *Proceedings of the 18th BioNLP Workshop and Shared*  
310 *Task*, pages 528–532.
- 311 Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and  
312 Balaraman Ravindran. 2020. Towards Transparent and Explainable Attention Models. In *Proceedings of the*  
313 *58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, Online. Association for  
314 Computational Linguistics.
- 315 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions  
316 of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and*  
317 *data mining*, pages 1135–1144.
- 318 Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of*  
319 *the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational  
320 Linguistics.
- 321 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and  
322 Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural*  
323 *Information Processing Systems, NIPS' 17*, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- 324 Sarah Wiegrefe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on*  
325 *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*  
326 *Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.
- 327 Qimin Zhou and Hao Wu. 2018. Nlp at iest 2018: Bilstm-attention and lstm-attention via soft voting in emotion  
328 classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and*  
329 *Social Media Analysis*, pages 189–194.

Model configuration	Single input sequence tasks						Dual input sequence tasks					
	SST	IMDB	Yelp	Amazon	20News	Tweets	SNLI	QQP	Babi1	Babi2	Babi3	CNN
Model LSTM variation	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla	Vanilla
Attention type	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh	tanh
Embedding dim	300	300	300	300	300	300	300	300	50	50	50	300
Embedding voc	13.826	12.487	63.328	49.883	6.515	6.845	20.981	26.635	24	38	39	70.190
Pre-embed	FastText	FastText	FastText	FastText	FastText	FastText	GLOVE	GLOVE	None	None	None	FastText
LSTM hidden dim	256	256	256	256	256	256	256	256	64	128	128	256
Output size	1	1	1	1	1	1	3	2	36 (6)*	36 (6)*	36 (6)*	584 (26)*
<b>Optimizer hyperparameters</b>												
Diversity weight (if appli)	0.5	0.5	0.5	0.5	0.5	0.5	0.1	0.5	0.5	0.5	0.5	0.2
Batch size	32	32	32	32	32	32	128	128	32	64	64	90
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
LR	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Weight decay	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05	1,E-05
Epochs	8	8	8	8	8	8	25	25	100	200	200	12
<b>Trainable weights, including fine-tuning of embeddings</b>												
Pencoder	4.719.192	4.317.492	19.569.792	15.536.292	2.525.892	2.624.892	6.865.692	8.480.892	30.896	94.060	94.110	21.628.392
Qencoder	0	0	0	0	0	0	6.865.692	8.480.892	30.896	94.060	94.110	21.628.392
Decoder	33.281	33.281	33.281	33.281	33.281	33.281	132.099	131.970	9.540	35.428	35.428	207.048
Total	4.752.473	4.350.773	19.603.073	15.569.573	2.559.173	2.658.173	13.863.483	17.093.754	71.332	223.548	223.648	43.463.832
<b>Trainable weights, without fine-tuning of embeddings</b>												
Pencoder	571.392	571.392	571.392	571.392	571.392	571.392	571.392	571.392	n/a	n/a	n/a	571.392
Qencoder	0	0	0	0	0	0	571.392	571.392	n/a	n/a	n/a	571.392
Decoder	33.281	33.281	33.281	33.281	33.281	33.281	132.099	131.970	n/a	n/a	n/a	207.048
Total	604.673	604.673	604.673	604.673	604.673	604.673	1.274.883	1.274.754	n/a	n/a	n/a	1.349.832

\* Output size is number of total entities in the dataset, part of which is masked in each datapoint (number of categories used on average per data point)

Table 8: Model- and hyperparameters for standard configurations per dataset

Dataset	Description	Number of datapoints			Avg seq. length (train)		Avg.no.answer categories	Vocab. size (train, docs)
		train (%pos)	dev (%pos)	test (%pos)	Document	Question		
<b>Single input sequence tasks</b>								
SST	Sentiment analysis	6,355 (52%)	821 (52%)	1,725 (50%)	21	n/a	2	13,703
IMDB	Sentiment analysis	17,200 (50%)	4,297 (49%)	4,356 (50%)	182	n/a	2	12,486
Yelp	Sentiment analysis	345,285 (54%)	4,790 (54%)	26,866 (54%)	74	n/a	2	63,304
Amazon*	Sentiment analysis	1,528,080 (52%)	4,456 (52%)	331,774 (52%)	57	n/a	2	49,881
Anemia*	Diagnosis prediction	-	-	-	-	-	-	-
Diabetes*	Diagnosis prediction	-	-	-	-	-	-	-
20News	Topic classification	1,145 (50%)	278 (50%)	357 (50%)	119	n/a	2	5,904
Tweets	Topic classification	13,938 (12%)	2,447 (13%)	4,123 (12%)	23	n/a	2	6,841
<b>Dual input sequence tasks</b>								
SNLI	Natural language inference	549,367	9,842	9,824	16	10	3	17,943
QQP	Paraphrase detection	327,460 (37%)	36,384 (37%)	40,430 (37%)	15	15	2	26,172
bAbI1	Question answering	8,500	1,500	1,000	38	5	6	20
bAbI2	Question answering	8,500	1,500	1,000	96	6	6	34
bAbI3	Question answering	8,500	1,500	1,000	309	9	6	34
CNN	Question answering	380,298	3,924	3,198	764	14	26.1	>70,000

\* Replication could not be performed for these datasets due to either availability or memory size limits

Table 9: Characteristics of the datasets

## 331 Appendix B: Full replication results

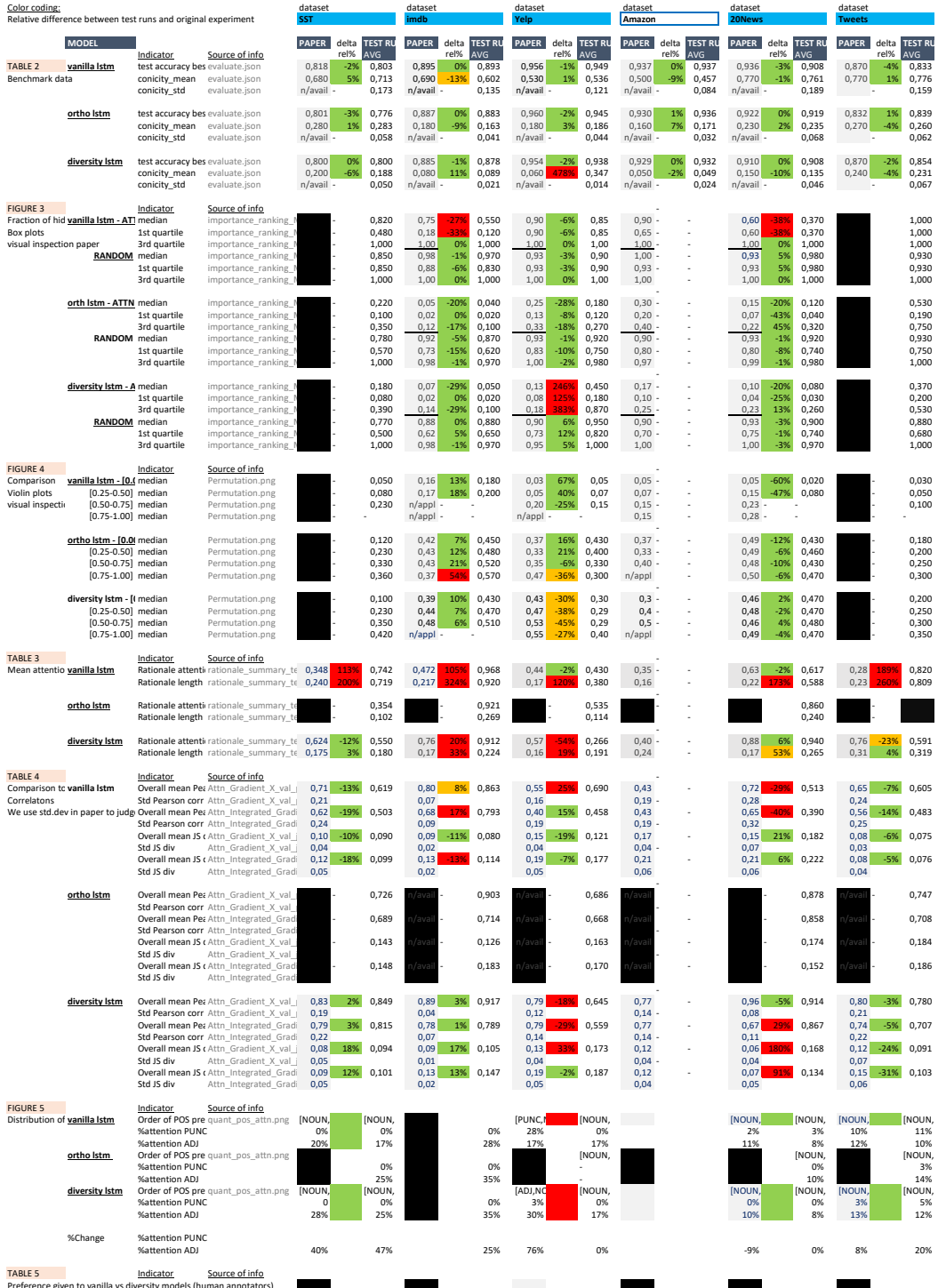


Figure 2: Replication results for single sequence tasks



### 332 Appendix C: Results of BiLSTM extension

Color coding:  
Relative difference vs. Vanilla Bi-LSTM  
(green = expected improvement)

MODEL	Indicator	Source of info	dataset			dataset			dataset			dataset		
			imdb			20News			QQP			Babi-1		
			PAPER	delta rel%	TEST RU AVG	PAPER	delta rel%	TEST RU AVG	PAPER	delta rel%	TEST RU AVG	PAPER	delta rel%	TEST RU AVG
TABLE 2 Benchmark data	vanilla lstm	test accuracy bes	0,895		0,893	0,936		0,908	0,787		0,784	0,991		1,000
		conicity_mean	0,690		0,602	0,770		0,761	0,590		0,583	0,560		0,766
		conicity_std	n/avail		0,135	n/avail		0,189	n/avail		0,131	n/avail		0,023
bi lstm	test accuracy bes	evaluate.json	-		0,905	-		0,919	-		0,790	-		0,997
	conicity_mean	evaluate.json	-		0,590	-		0,818	-		0,610	-		0,692
	conicity_std	evaluate.json	-		0,118	-		0,136	-		0,103	-		0,042
ortho bi lstm	test accuracy bes	evaluate.json	-1%		0,893	-4%		0,885	-1%		0,785	0%		1,000
	conicity_mean	evaluate.json	-70%		0,176	-54%		0,373	-44%		0,341	-68%		0,220
	conicity_std	evaluate.json	-		0,045	-		0,156	-		0,068	-		0,055
diversity bi lstm	test accuracy bes	evaluate.json	-1%		0,892	1%		0,924	-1%		0,786	0%		1,000
	conicity_mean	evaluate.json	-76%		0,144	-80%		0,164	-94%		0,035	-93%		0,046
	conicity_std	evaluate.json	-		0,025	-		0,074	-		0,019	-		0,016
FIGURE 3	bi lstm - ATTN	median			0,980			0,520			0,830			0,030
		1st quartile			0,670			0,520			0,380			0,020
		3rd quartile			1,000			1,000			1,000			0,080
	RANDOM	median			0,920			0,980			0,740			0,600
		1st quartile			0,920			0,980			0,740			0,320
		3rd quartile			1,000			1,000			1,000			0,820
	ortho bi lstm - A1	median		-93%	0,070	-52%	0,250	-19%	0,670	-33%	0,020			
		1st quartile		-97%	0,020	-92%	0,040	-39%	0,230	-50%	0,010			
		3rd quartile		-87%	0,130	-61%	0,390	0%	1,000	-50%	0,040			
	RANDOM	median		0%	0,920	-3%	0,950	-15%	0,630	-13%	0,520			
		1st quartile		-22%	0,720	-17%	0,810	-15%	0,630	-16%	0,270			
		3rd quartile		-3%	0,970	0%	1,000	0%	1,000	-7%	0,760			
	diversity bi lstm	median		-93%	0,070	-88%	0,060	-40%	0,500	-33%	0,020			
		1st quartile		-97%	0,020	-94%	0,030	32%	0,500	-50%	0,010			
		3rd quartile		-67%	0,330	-87%	0,130	0%	1,000	-50%	0,040			
	RANDOM	median		0%	0,920	-8%	0,900	-1%	0,730	-13%	0,520			
		1st quartile		-24%	0,700	-27%	0,720	-1%	0,730	-16%	0,270			
		3rd quartile		-2%	0,980	-2%	0,980	0%	1,000	-6%	0,770			
FIGURE 4	Comparison Violin plots visual inspectic	vanilla lstm - [0.0	Permutation.png	0,16	0,180	0,05	0,020			0,070				
		0.25-0.50]	Permutation.png	0,17	0,200	0,15	0,080			0,100			0,820	
		0.50-0.75]	Permutation.png	n/appl		0,23				0,100			0,870	
	bi lstm - [0.00-0.2	0.25-0.50]	Permutation.png		0,100		0,010			0,020				
		0.50-0.75]	Permutation.png		0,090		0,200			0,080			0,570	
		0.75-1.00]	Permutation.png				0,270			0,210			0,920	
	ortho bi lstm - [0	0.25-0.50]	Permutation.png				0,730			0,430			0,970	
		0.50-0.75]	Permutation.png											
		0.75-1.00]	Permutation.png											
	diversity bi lstm	0.25-0.50]	Permutation.png											
		0.50-0.75]	Permutation.png											
		0.75-1.00]	Permutation.png											
	TABLE 4	Comparison to vanilla lstm	Overall mean Pea Attn_Gradient_X_val_j	0,80	0,863	0,72	0,513	0,19	0,410	0,56	0,959			
			Overall mean Pea Attn_Integrated_Gradi	0,68	0,793	0,65	0,390	-0,06	0,084	0,33	0,785			
			Overall mean JS c Attn_Gradient_X_val_j	0,09	0,080	0,15	0,182	0,15	0,101	0,33	0,120			
		bi lstm	Overall mean JS c Attn_Integrated_Gradi	0,13	0,114	0,21	0,222	0,19	0,148	0,43	0,247			
			Overall mean Pea Attn_Gradient_X_val_j		0,817		0,535		0,345		0,867			
			Overall mean Pea Attn_Integrated_Gradi		0,680		0,411		0,018		0,643			
ortho bi lstm		Overall mean JS c Attn_Gradient_X_val_j		0,091		0,116		0,155		0,210				
		Overall mean JS c Attn_Integrated_Gradi		0,147		0,141		0,213		0,343				
		Overall mean Pea Attn_Gradient_X_val_j	14%	0,930	-35%	0,346	32%	0,457	-1%	0,862				
diversity bi lstm		Overall mean Pea Attn_Integrated_Gradi	25%	0,849	-4%	0,394	1894%	0,359	15%	0,738				
		Overall mean JS c Attn_Gradient_X_val_j	-5%	0,086	258%	0,415	-14%	0,133	-5%	0,199				
		Overall mean JS c Attn_Integrated_Gradi	-29%	0,105	174%	0,386	-31%	0,148	-15%	0,292				
Overall mean Pea Attn_Gradient_X_val_j		11%	0,908	79%	0,960	50%	0,516	-65%	0,307					
		Overall mean Pea Attn_Integrated_Gradi	25%	0,847	114%	0,880	1756%	0,334	-11%	0,571				
		Overall mean JS c Attn_Gradient_X_val_j	-19%	0,108	-16%	0,097	-34%	0,103	106%	0,433				
Overall mean JS c Attn_Integrated_Gradi		-16%	0,124	-23%	0,108	-39%	0,130	5%	0,359					

Figure 4: Performance of BiLSTM on evaluation metrics

**SST examples**

SST, Vanilla-LSTM, trained on default (tanh) attention (test acc=**.805**, conicity=**.697**)

Sentence	: <SOS>	a	slick	skillful	little	horror	film	<EOS>
Attentions	: 0.00	0.04	0.08	<b>0.20</b>	<b>0.30</b>	<b>0.24</b>	0.13	0.00
SM( h <sub>i</sub>  )	: 0.00	0.09	0.17	0.24	0.17	0.17	0.17	0.00
Label:	1	, Prediction: [0.5812621]						

SST, Ortho-LSTM, trained on default (tanh) attention (test acc=**.776**, conicity=**.283**)

Sentence	: <SOS>	a	slick	skillful	little	horror	film	<EOS>
Attentions	: 0.00	0.00	0.03	<b>0.64</b>	0.18	0.04	0.11	0.00
SM( h <sub>i</sub>  )	: 0.00	0.15	0.14	0.21	0.10	0.16	0.25	0.00
Label:	1	, Prediction: [0.7968075]						

SST, Diversity-LSTM, trained on default (tanh) attention (test acc=**.800**, conicity=**.188**)

Sentence	: <SOS>	a	slick	skillful	little	horror	film	<EOS>
Attentions	: 0.00	0.00	0.12	<b>0.84</b>	0.01	0.01	0.01	0.00
SM( h <sub>i</sub>  )	: 0.00	0.02	0.04	0.19	0.25	0.16	0.34	0.00
Label:	1	, Prediction: [0.9623399]						

**Yelp examples**

Yelp, Vanilla-LSTM, trained on default (tanh) attention (test acc=**.949**, conicity=**.536**)

Sentence	: <SOS>	Been	going	here	for	years.	A	great	place!	<EOS>
Attentions	: 0.00	<b>0.35</b>	0.09	0.06	0.01	0.01	0.04	0.18	<b>0.24</b>	0.00
SM( h <sub>i</sub>  )	: 0.00	0.08	0.07	0.10	0.08	0.09	0.06	0.18	0.35	0.00
Label:	1	, Prediction: [0.984035]								

Yelp, Ortho-LSTM, trained on default (tanh) attention (test acc=**.945**, conicity=**.186**)

Sentence	: <SOS>	Been	going	here	for	years.	A	great	place!	<EOS>
Attentions	: 0.00	<b>0.53</b>	0.00	0.00	0.00	0.08	0.00	<b>0.22</b>	0.17	0.00
SM( h <sub>i</sub>  )	: 0.00	0.18	0.08	0.05	0.07	0.04	0.03	0.34	0.21	0.00
Label:	1	, Prediction: [0.99045765]								

Yelp, Diversity-LSTM, trained on default (tanh) attention (test acc=**.938**, conicity=**.347**)

Sentence	: <SOS>	Been	going	here	for	years.	A	great	place!	<EOS>
Attentions	: 0.00	<b>0.41</b>	0.16	0.00	0.00	0.02	0.09	<b>0.21</b>	0.10	0.00
SM( h <sub>i</sub>  )	: 0.00	0.18	0.05	0.02	0.04	0.03	0.06	0.40	0.21	0.00
Label:	1	, Prediction: [0.9969946]								

Figure 5: Examples of single input sequence tasks

## CNN examples

CNN, Vanilla-LSTM, trained with default (tanh) attention (test acc=.595, conicity=.395)

P path : <SOS> ( @entity2 ) one @entity1 citizen was killed and another injured in what  
 Attentions : 0.00 0.00 0.00 0.00 0.04 0.06 0.08 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 police are calling a suspected terror attack wednesday night near @entity6 . @entity8  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.07  
 spokesman @entity7 said a 37 - year - old @entity10 motorist from @entity11 struck two  
 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.43 0.00 0.00 0.03 0.00 0.00 0.00  
 people standing at a bus stop in the @entity15 section of the city . one victim ,  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 identified by police as @entity18 , 26 , died at the hospital . a 20 - year -  
 0.00 0.00 0.00 0.00 0.20 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 old woman remains in serious condition , according to @entity7 . the driver has been  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 arrested and is under investigation by the @entity24 . " from the investigation and first  
 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 findings , there is a strong suspicion that we 're talking about a terror attack ,  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 " @entity7 said . amid the ongoing investigation , a magistrate court has issued a  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 gag order on details of the incident . <EOS>  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00

Q path : <SOS> the suspect is a 37 - year - old @placeholder from @entity11 , @entity1 police say <EOS>

Answer: @entity10 Predicted: @entity10

CNN, Ortho-LSTM, trained with default (tanh) attention (test acc=.536, conicity=.099)

P path : <SOS> ( @entity2 ) one @entity1 citizen was killed and another injured in what  
 Attentions : 0.00 0.00 0.15 0.00 0.00 0.00 0.00 0.01 0.00 0.02 0.00 0.00 0.02 0.00  
 police are calling a suspected terror attack wednesday night near @entity6 . @entity8  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.03 0.00  
 spokesman @entity7 said a 37 - year - old @entity10 motorist from @entity11 struck two  
 0.00 0.00 0.03 0.02 0.00 0.02 0.00 0.01 0.00 0.00 0.00 0.00 0.01 0.00  
 people standing at a bus stop in the @entity15 section of the city . one victim ,  
 0.00 0.05 0.13 0.06 0.00 0.00 0.15 0.01 0.00 0.04 0.03 0.01 0.00 0.02 0.00 0.00  
 identified by police as @entity18 , 26 , died at the hospital . a 20 - year -  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.01 0.00 0.02 0.00 0.01  
 old woman remains in serious condition , according to @entity7 . the driver has been  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00  
 arrested and is under investigation by the @entity24 . " from the investigation and first  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 findings there is a strong suspicion that we 're talking about a terror attack ,  
 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 " @entity7 said . amid the ongoing investigation , a magistrate court has issued a  
 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  
 gag order on details of the incident . <EOS>  
 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00

Q path : <SOS> the suspect is a 37 - year - old @placeholder from @entity11 , @entity1 police say <EOS>

Answer: @entity10 Predicted: @entity10

Figure 6: Examples of dual input sequence tasks