Attention is not Explanation

Mark Jacobsen

Assumptions and Research Questions

Questions:
Does attention provide model transparency?
Are attended-to features responsible for outputs?

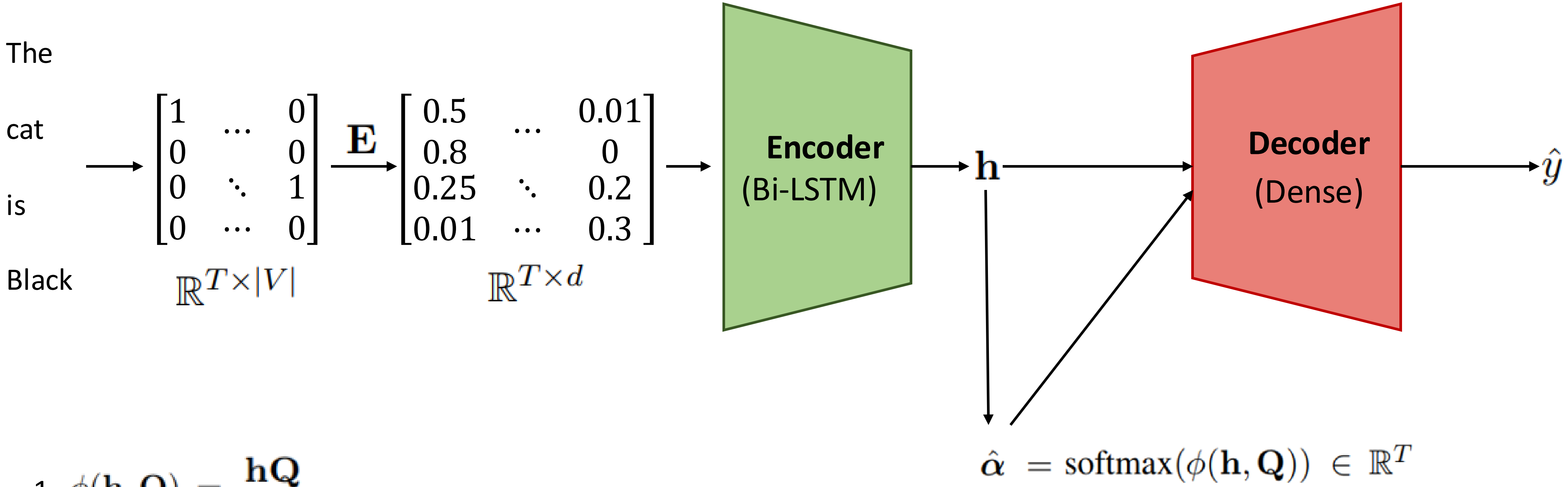
Authors claim: No

If yes:
1. Attention weights should correlate with feature importance methods.
   - Gradient-based methods
   - Leave-one-out

2. Alternative attention weight configurations should yield corresponding changes in prediction.

Experiments

Attention / Model Architecture



1. $\phi(\mathbf{h}, \mathbf{Q}) = \dfrac{\mathbf{hQ}}{\sqrt{m}}$

2. $\phi(\mathbf{h}, \mathbf{Q}) = \mathbf{v}^T \tanh(\mathbf{W_1 h} + \mathbf{W_2 Q})$

Experiments

Mark Jacobsen

Datasets / NLP Tasks

Natural Language Inference:
- SNLI Dataset

Sentiment Analysis:
- Stanford Sentiment Treebank (SST)
- IMDB Large Movie Revies Corpus

Other Binary Text Classification:
- Twitter Adverse Drug Reaction dataset
- 20 Newsgroups (Hockey vs Basketball)
- AG News Corpus
- …

Question Answering:
- CNN News Articles
- bAbI

Dataset	$\|V\|$	*Avg. length*	*Train size*	*Test size*	*Test performance*
SST	16175	19	3034 / 3321	863 / 862	0.81
IMDB	13916	179	12500 / 12500	2184 / 2172	0.88
ADR Tweets	8686	20	14446 / 1939	3636 / 487	0.61
20 Newsgroups	8853	115	716 / 710	151 / 183	0.94
AG News	14752	36	30000 / 30000	1900 / 1900	0.96
Diabetes (MIMIC)	22316	1858	6381 / 1353	1295 / 319	0.79
Anemia (MIMIC)	19743	2188	1847 / 3251	460 / 802	0.92
CNN	74790	761	380298	3198	0.64
bAbI (Task 1 / 2 / 3)	40	8 / 67 / 421	10000	1000	1.0 / 0.65 / 0.64
SNLI	20982	14	182764 / 183187 / 183416	3219 / 3237 / 3368	0.78

Feature Importance Correlation

1. Attention weights should correlate with feature importance methods.

Algorithm 1 Feature Importance Computations

$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x}),\ \hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$

$\hat{y} \leftarrow \text{Dec}(\mathbf{h}, \alpha)$

Model Computation

$g_t \leftarrow |\sum_{w=1}^{|V|} \mathbb{1}[\mathbf{x}_{tw} = 1] \frac{\partial y}{\partial \mathbf{x}_{tw}}|\ , \forall t \in [1, T]$

$\tau_g \leftarrow \text{Kendall-}\tau(\alpha, g)$

Gradient Based Feature Importance Kendall Correlation

$\Delta \hat{y}_t \leftarrow \text{TVD}(\hat{y}(\mathbf{x}_{-t}), \hat{y}(\mathbf{x}))\ , \forall t \in [1, T]$

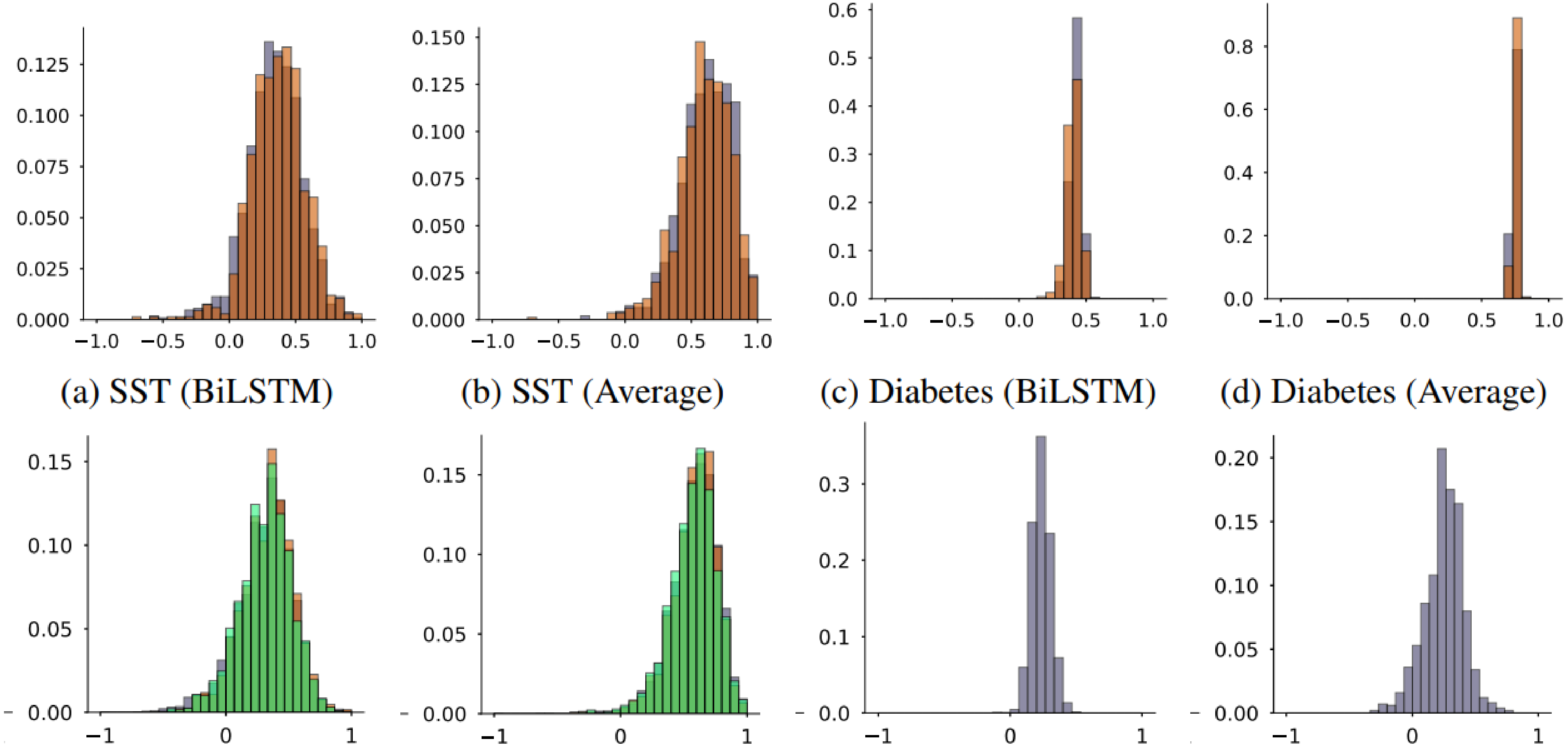
$\tau_{loo} \leftarrow \text{Kendall-}\tau(\alpha, \Delta \hat{y})$

Leave-One-Out Feature Importance Kendall Correlation

Feature Importance Correlation

		Gradient (BiLSTM) τ_g		Gradient (Average) τ_g		Leave-One-Out (BiLSTM) τ_{loo}	
Dataset	Class	Mean $\pm$ Std.	Sig. Frac.	Mean $\pm$ Std.	Sig. Frac.	Mean $\pm$ Std.	Sig. Frac.
SST	0	0.34 ± 0.21	0.48	0.61 ± 0.20	0.87	0.27 ± 0.19	0.33
	1	0.36 ± 0.21	0.49	0.60 ± 0.21	0.83	0.32 ± 0.19	0.40
IMDB	0	0.44 ± 0.06	1.00	0.67 ± 0.05	1.00	0.34 ± 0.07	1.00
	1	0.43 ± 0.06	1.00	0.68 ± 0.05	1.00	0.34 ± 0.07	0.99
ADR Tweets	0	0.47 ± 0.18	0.76	0.73 ± 0.13	0.96	0.29 ± 0.20	0.44
	1	0.49 ± 0.15	0.85	0.72 ± 0.12	0.97	0.44 ± 0.16	0.74
20News	0	0.07 ± 0.17	0.37	0.79 ± 0.07	1.00	0.06 ± 0.15	0.29
	1	0.21 ± 0.22	0.61	0.75 ± 0.08	1.00	0.20 ± 0.20	0.62
AG News	0	0.36 ± 0.13	0.82	0.78 ± 0.07	1.00	0.30 ± 0.13	0.69
	1	0.42 ± 0.13	0.90	0.76 ± 0.07	1.00	0.43 ± 0.14	0.91
Diabetes	0	0.42 ± 0.05	1.00	0.75 ± 0.02	1.00	0.41 ± 0.05	1.00
	1	0.40 ± 0.05	1.00	0.75 ± 0.02	1.00	0.45 ± 0.05	1.00
Anemia	0	0.47 ± 0.05	1.00	0.77 ± 0.02	1.00	0.46 ± 0.05	1.00
	1	0.46 ± 0.06	1.00	0.77 ± 0.03	1.00	0.47 ± 0.06	1.00
CNN	Overall	0.24 ± 0.07	0.99	0.50 ± 0.10	1.00	0.20 ± 0.07	0.98
bAbI 1	Overall	0.25 ± 0.16	0.55	0.72 ± 0.12	0.99	0.16 ± 0.14	0.28
bAbI 2	Overall	-0.02 ± 0.14	0.27	0.68 ± 0.06	1.00	-0.01 ± 0.13	0.27
bAbI 3	Overall	0.24 ± 0.11	0.87	0.61 ± 0.13	1.00	0.26 ± 0.10	0.89
SNLI	0	0.31 ± 0.23	0.36	0.59 ± 0.18	0.80	0.16 ± 0.26	0.20
	1	0.33 ± 0.21	0.38	0.58 ± 0.19	0.80	0.36 ± 0.19	0.44
	2	0.31 ± 0.21	0.36	0.57 ± 0.19	0.80	0.34 ± 0.20	0.40

Gradient Feature Importance Correlation



(a) SST (BiLSTM) (b) SST (Average) (c) Diabetes (BiLSTM) (d) Diabetes (Average)

(e) SNLI (BiLSTM) (f) SNLI (Average) (g) CNN-QA (BiLSTM) (h) BAbI 1 (BiLSTM)

Attention Changes

2. Alternative attention weight configurations should yield corresponding changes in prediction.

Algorithm 2 Permuting attention weights

$\mathbf{h} \leftarrow \mathrm{Enc}(\mathbf{x}),\ \hat{\alpha} \leftarrow \mathrm{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \mathrm{Dec}(\mathbf{h}, \hat{\alpha})$

Model Computation

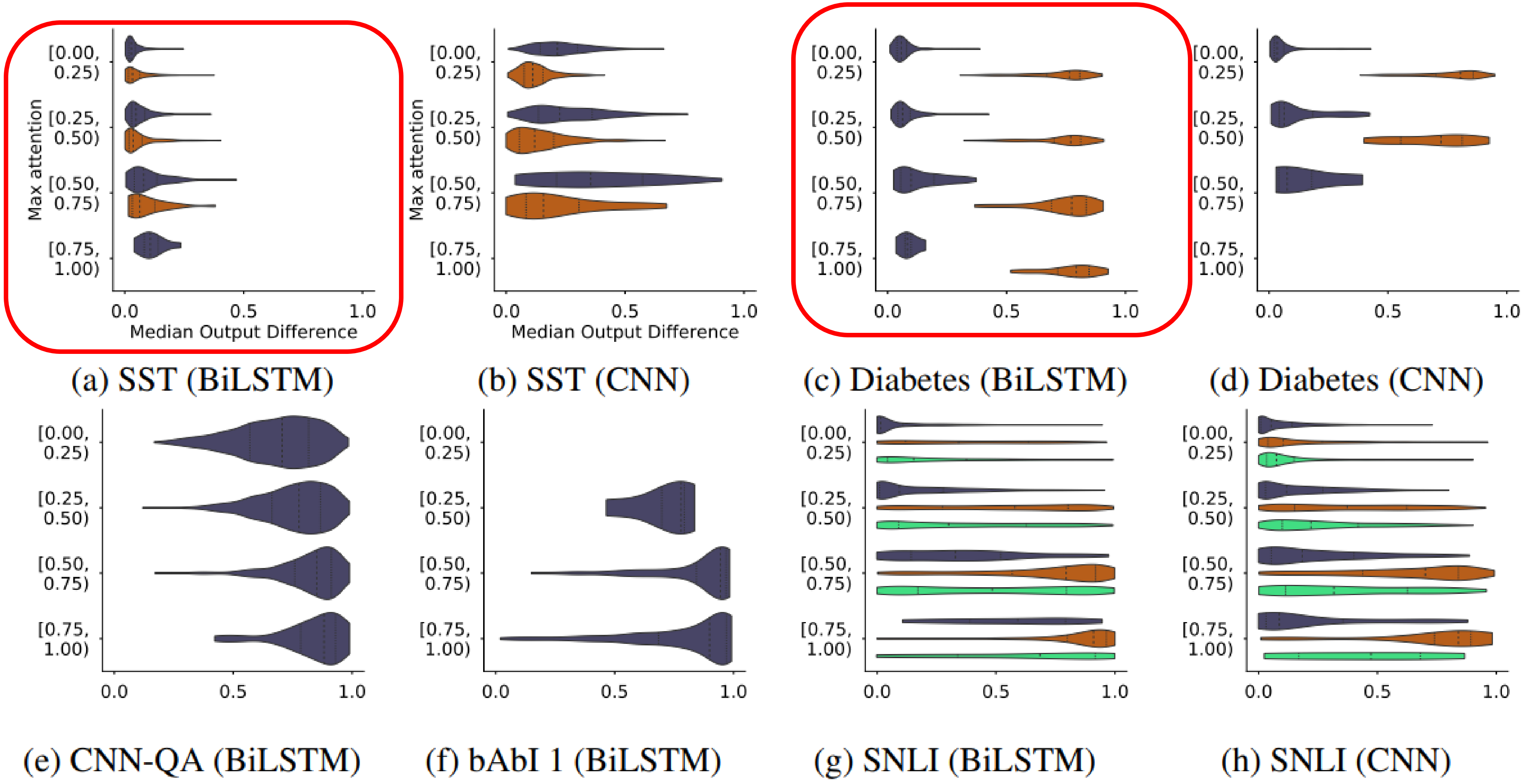
for $p \leftarrow 1$ to 100 **do**
  $\alpha^p \leftarrow \mathrm{Permute}(\hat{\alpha})$
  $\hat{y}^p \leftarrow \mathrm{Dec}(\mathbf{h}, \alpha^p)$ ▷ Note : $\mathbf{h}$ is not changed
  $\Delta\hat{y}^p \leftarrow \mathrm{TVD}[\hat{y}^p, \hat{y}]$
end for

Random Attention Permutation
Compute Output Difference

$\Delta\hat{y}^{med} \leftarrow \mathrm{Median}_p(\Delta\hat{y}^p)$

Median Output Difference

Random Attention Permutation



(a) SST (BiLSTM) (b) SST (CNN) (c) Diabetes (BiLSTM) (d) Diabetes (CNN)

(e) CNN-QA (BiLSTM) (f) bAbI 1 (BiLSTM) (g) SNLI (BiLSTM) (h) SNLI (CNN)

Adversarial Attention

2. Alternative attention weight configurations should yield corresponding changes in prediction.

Algorithm 3 Finding adversarial attention weights

$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x}),\ \hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \text{Dec}(\mathbf{h}, \hat{\alpha})$

Model Computation

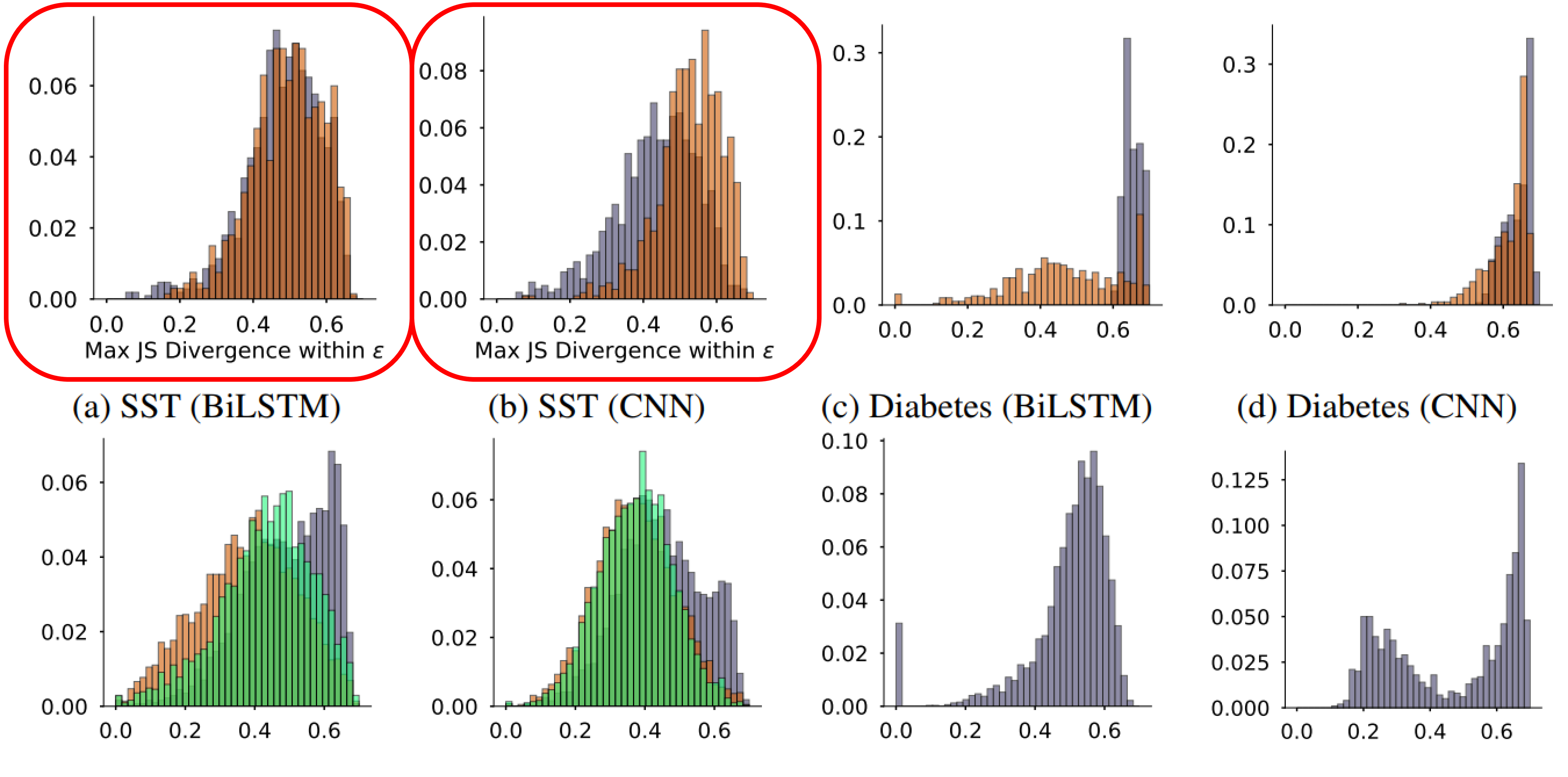
$\alpha^{(1)}, ..., \alpha^{(k)} \leftarrow$ Optimize Eq 1
for $i \leftarrow 1$ to k **do**
  $\hat{y}^{(i)} \leftarrow \text{Dec}(\mathbf{h}, \alpha^{(i)})$ ▷ $\mathbf{h}$ is not changed
  $\Delta\hat{y}^{(i)} \leftarrow \text{TVD}[\hat{y}, \hat{y}^{(i)}]$
  $\Delta\alpha^{(i)} \leftarrow \text{JSD}[\hat{\alpha}, \alpha^{(i)}]$
end for

Adversarially Optimize Attention
Compute difference in
- Output
- Attention Weight

$\epsilon\text{-max JSD} \leftarrow \max_i \mathbb{1}[\Delta\hat{y}^{(i)} \leq \epsilon]\Delta\alpha^{(i)}$

Find maximum attention difference with maximum threshold output difference

Adversarial Attention



(a) SST (BiLSTM) (b) SST (CNN) (c) Diabetes (BiLSTM) (d) Diabetes (CNN)

(e) SNLI (BiLSTM) (f) SNLI (CNN) (g) CNN-QA (BiLSTM) (h) BAbI 1 (BiLSTM)

Results

Prior Assumptions

1. Attention weights should correlate with feature importance methods.

- Gradient-based methods
- Leave-one-out

Weak correlation

2. Alternative attention weight configurations should yield corresponding changes in prediction.

Negligible Changes

→ Attention is Not Explanation

Discussion

Mark Jacobsen

Pros / Cons

Positive:

- Clear research question and experiments for evaluation
- Various datasets and empirical evidence to backup claims
- They use previously established definitions of “explainability” / “transparency”
    - Still up to debate

Negative:

- Questionable Assumptions
    - Using feature attribution as ground-truth
    - Explanations need to be exclusive
- Strong focus on binary classification and LSTMs
- Is changing the attention weights valid?
- Their own results sometimes show that attention can be explanation

Impact / Debate

1661 Citations

2019
Attention is not Explanation

2019
Attention is not not Explanation

2019
Is Attention Interpretable?

2020
(How) Can Attention Become Explanation?

2020
Why is Attention Not So Interpretable

2021
Why Attention May Not Be Interpretable?

2021
Is Sparse Attention more Interpretable?

2022
Is Attention Explanation? An Introduction to the Debate

2022
Attention cannot be an Explanation

2023
SEAT: Stable and Explainable Attention

2024
On Explaining with Attention Matrices

Sources

[1] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. arXiv preprint arXiv:1902.10186. Retrieved from https://arxiv.org/abs/1902.10186.

[2] Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. arXiv preprint arXiv:1908.04626. Retrieved from https://arxiv.org/abs/1908.04626.

[3] Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., & Watrin, P. (2022). Is attention explanation? An introduction to the debate. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3889–3900). Association for Computational Linguistics. https://aclanthology.org/2022.acl-long.269/.