Attention is not Explanation

Authors: Sarthak Jain, Byron C. Wallace
Presenters: Nikhil Nayak, Sree Harsha Tanneru, Hongjin Lin

2023/03/06

Attention Mechanism is a “recent” breakthrough in NLP

2015: birth of attention mechanism in NLP

- “Neural machine translation by jointly learning to align and translate” Bahdanau et al. 2015

2017: birth of transformers based on self attention mechanism

- “Attention is all you need” Vaswani et al. 2017

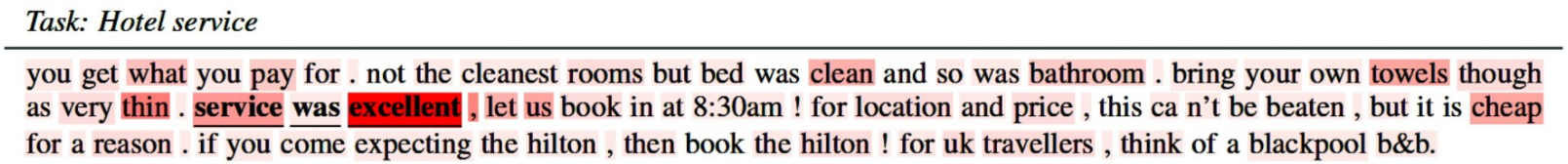


Fig. 1. Example of attention visualization for an aspect-based sentiment analysis task, from [1, Fig. 6]. Words are highlighted according to attention scores. Phrases in bold are the words considered relevant for the task or human rationales.

Source: Galassi et al. 2021

Many use attention as an explanation mechanism

Show, Attend
Genera

Kelvin Xu
Jimmy Lei Ba
Ryan Kiros
Kyunghyun Cho
Aaron Courville
Ruslan Salakhutdinov
Richard S. Zemel
Yoshua Bengio

RETA
Healthca

Chen *et al. BMC Medical Informatics and Decision Making* 2020, **20**(Suppl 3):131
https://doi.org/10.1186/s12911-020-1110-7

BMC Medical Informatics and Decision Making

RESEARCH Open Access

Interpretable clinical prediction via attention-based neural network

Peipei Chen[1,2], Wei Dong[3], Jinliang Wang[4], Xudong Lu[1,2], Uzay Kaymak[2,1] and Zhengxing Huang[1*]

From 5th China Health Information Processing Conference
Guangzhou, China. 22-24 November 2019

Andy Schuetz†, Walter F. Stewart†, Jimeng Sun*
* Georgia Institute of Technology † Sutter Health
{mp2893,bahadori,jkulas3}@gatech.edu,
{schueta1,stewarwf}@sutterhealth.org, jsun@cc.gatech.edu

But does attention necessarily provide faithful explanation?

What do you think?

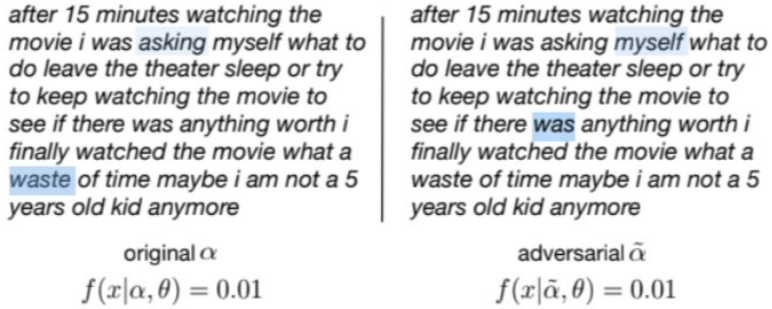


Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

Research Questions

What is the degree to which attention weights provide meaningful “explanations" for predictions?

1. Consistency: To what extent do induced attention weights **correlate with measures of feature importance** – specifically, those resulting from gradients and leave-one-out (LOO) methods?

2. Counterfactual attention weights: Would **alternative attention weights** (and hence distinct heatmaps/“explanations”) necessarily yield different predictions?

Contributions and main findings

1. Standard attention modules **do not provide meaningful explanations** and should not be treated as though they do.

2. Learned attention weights are frequently **uncorrelated with gradient-based** measures of feature importance.

3. One can identify very **different attention distributions** that nonetheless yield equivalent predictions.

Experimental setup

Dataset	$\|V\|$	*Avg. length*	*Train size*	*Test size*	*Test performance (LSTM)*
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 / 0.48 / 0.62
SNLI	20982	14	182764 / 183187 / 183416	3219 / 3237 / 3368	0.78

Binary Text Classification

Question Answering

Natural Language Inference

1 - Correlation between attention and feature importance

2 - Counterfactual attention weights

Correlation between Attention and Feature Importance

To what extent do induced attention weights correlate with measures of feature importance?

Specifically two methods are considered:

1. Feature gradient methods
2. Leave-One-Out (LOO) method

Gradient based Feature Importance

- To use the feature gradient method, we compute the gradient of the output with respect to each input feature.

- This tells us how much changing each feature would affect the output.

- We can visualize this by highlighting the input features that have the highest absolute gradient values.

Feature Erasure/LOO

- To use the LOO method, we remove one input feature at a time and observe how the output changes.

- This allows us to measure the impact of each feature on the output.

- We can visualize this by highlighting the input feature that is removed and showing the change in the output.

Distribution Change Measure for Experiments

Total Variation Distance (TVD), Kendall's τ coefficient

$$\mathrm{TVD}(\hat{y}_1, \hat{y}_2) = \frac{1}{2} \sum_{i=1}^{|\mathcal{Y}|} |\hat{y}_{1i} - \hat{y}_{2i}|$$

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})} = 1 - \frac{2(\text{number of discordant pairs})}{\binom{n}{2}}$$

Algorithm for Feature Importance Computations

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

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

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

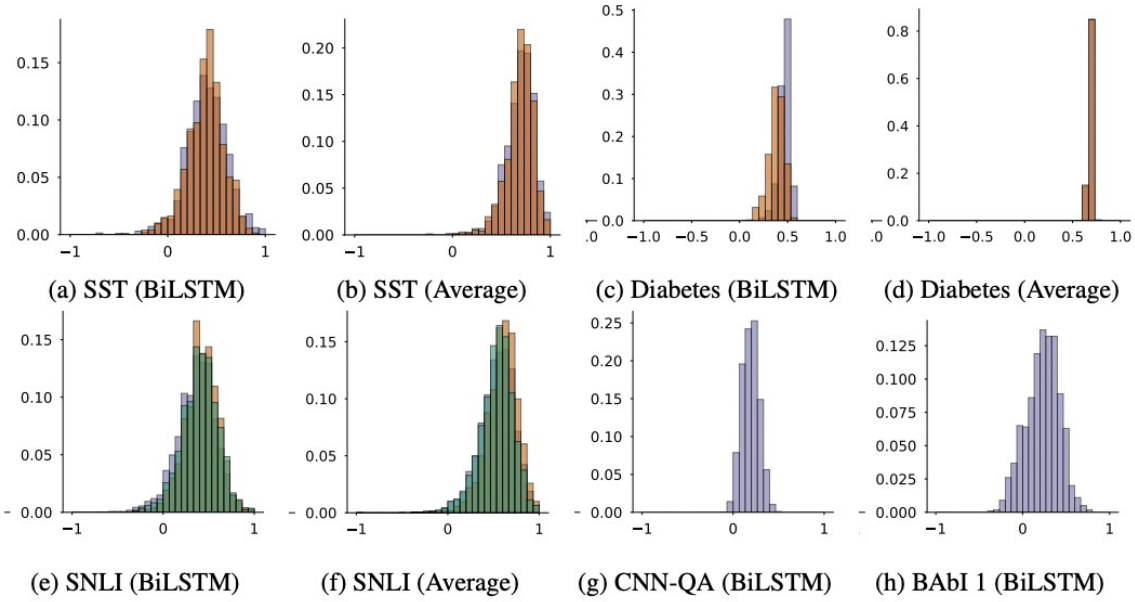
$$\tau_g \leftarrow \text{Kendall-}\tau(\alpha, g)$$

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

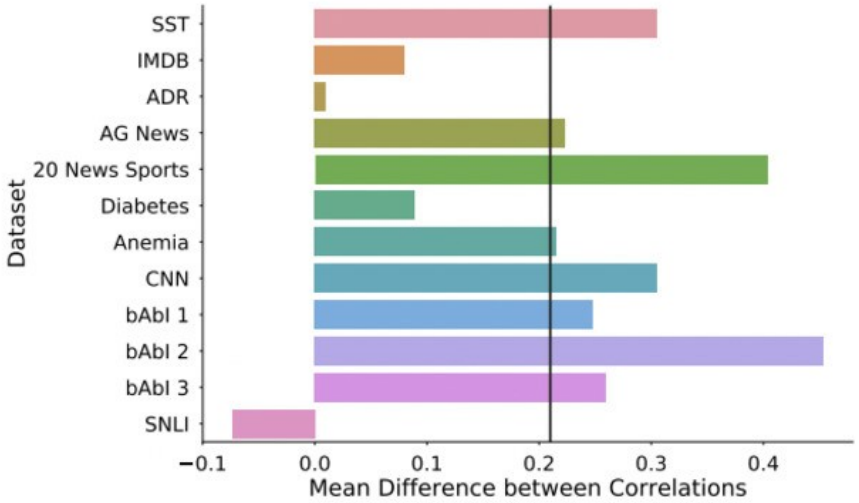
$$\tau_{loo} \leftarrow \text{Kendall-}\tau(\alpha, \Delta\hat{y})$$

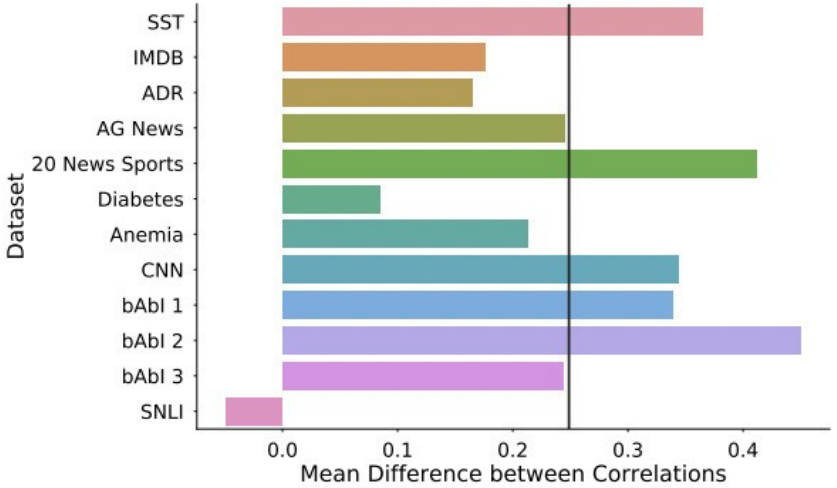
Feature Gradients Results

Histogram of Kendall correlation between attention and gradients.



Correlation Comparison - Attention, Gradients, LOO





Mean difference in correlation.

Fig 1 - LOO, Gradients V/S Attention, LOO. Fig 2 - LOO, Gradients V/S Attention, Gradients.

Possible Limitations in Experiment Results

1. Potential issues with Kendall T correlation measure - Many irrelevant features may add noise to the correlation measures.

2. Alternative measures of feature importance, such as gradients and LOO, may not necessarily be considered ideal or deemed as the 'ground truth' for comparison purposes.

Counterfactual Attention Weights

What is the degree to which attention weights provide meaningful “explanations" for predictions?

1. **Consistency** i.e; correlation with other measures of feature importance
2. **Counterfactual attention distributions** should yield corresponding changes in predictions.

Counterfactual Attention Weights

Would the prediction be different if the model attended to different input features ?

Two experiments

1. **Attention Permutation**: Scramble the original attention weights
2. **Adversarial Attention**: Generate an adversarial distribution that is maximally distinct from the observed attention weights, whilst simultaneously close to the observed prediction

Attention Permutation - Algorithm

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

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

for $p \leftarrow 1$ to 100 **do**

$$\alpha^p \leftarrow \text{Permute}(\hat{\alpha})$$

$$\hat{y}^p \leftarrow \text{Dec}(\mathbf{h}, \alpha^p) \qquad \triangleright \text{ Note : } \mathbf{h} \text{ is not changed}$$

$$\Delta\hat{y}^p \leftarrow \text{TVD}[\hat{y}^p, \hat{y}]$$

end for

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

Attention Permutation - Results

Plot Δy at different peak attention values

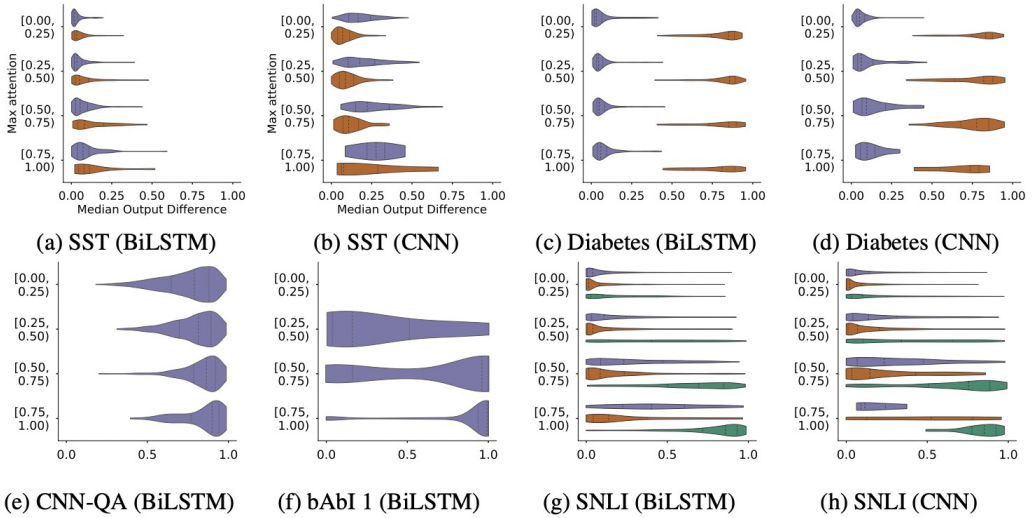


Figure 6: **Median change in output $\Delta\hat{y}^{med}$** (x-axis) densities in relation to the **max attention** ($\max \hat{\alpha}$) (y-axis) obtained by randomly permuting instance attention weights. Encoders denoted parenthetically. Plots for all corpora and using all encoders are available online.

Adversarial Attention

- Explicitly seek out attention weights that differ as much as possible from observed attention weights, yet leaving prediction almost unchanged.
- Why ?
- Alternative attention distributions identified for the same output may be viewed as equally plausible explanations for the same output. Is this undesirable ?
    - Cannot conclude that model made a specific prediction because it attended over inputs in a certain way.
    - Lack of specificity makes constructing counterfactual explanations hard ?
    - Makes model less trustworthy ?

Adversarial Attention - Optimisation

Seek $\alpha^{(1)}$, $\alpha^{(2)}$, $\alpha^{(3)}$ ... $\alpha^{(k)}$.

ϵ - Quantifies small difference in model output

Measure of distance between attention distributions ?

$$JSD(\alpha_1, \alpha_2) = \frac{1}{2}(KL[\alpha_1|\frac{\alpha_1 + \alpha_2}{2}] + KL[\alpha_2|\frac{\alpha_1 + \alpha_2}{2}])$$

$$\begin{aligned} \underset{\alpha^{(1)},\ldots,\alpha^{(k)}}{\text{maximize}} \quad & f(\{\alpha^{(i)}\}_{i=1}^{k}) \\ \text{subject to} \quad & \forall i \ \text{TVD}[\hat{y}(\mathbf{x}, \alpha^{(i)}), \hat{y}(\mathbf{x}, \hat{\alpha})] \leq \epsilon \end{aligned} \tag{1}$$

Where $f(\{\alpha^{(i)}\}_{i=1}^{k})$ is:

$$\sum_{i=1}^{k} \text{JSD}[\alpha^{(i)}, \hat{\alpha}] + \frac{1}{k(k-1)} \sum_{i<j} \text{JSD}[\alpha^{(i)}, \alpha^{(j)}] \tag{2}$$

Adversarial Attention - Algorithm

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

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

$\alpha^{(1)}, ..., \alpha^{(k)} \leftarrow$ 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

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

Adversarial Attention - Results

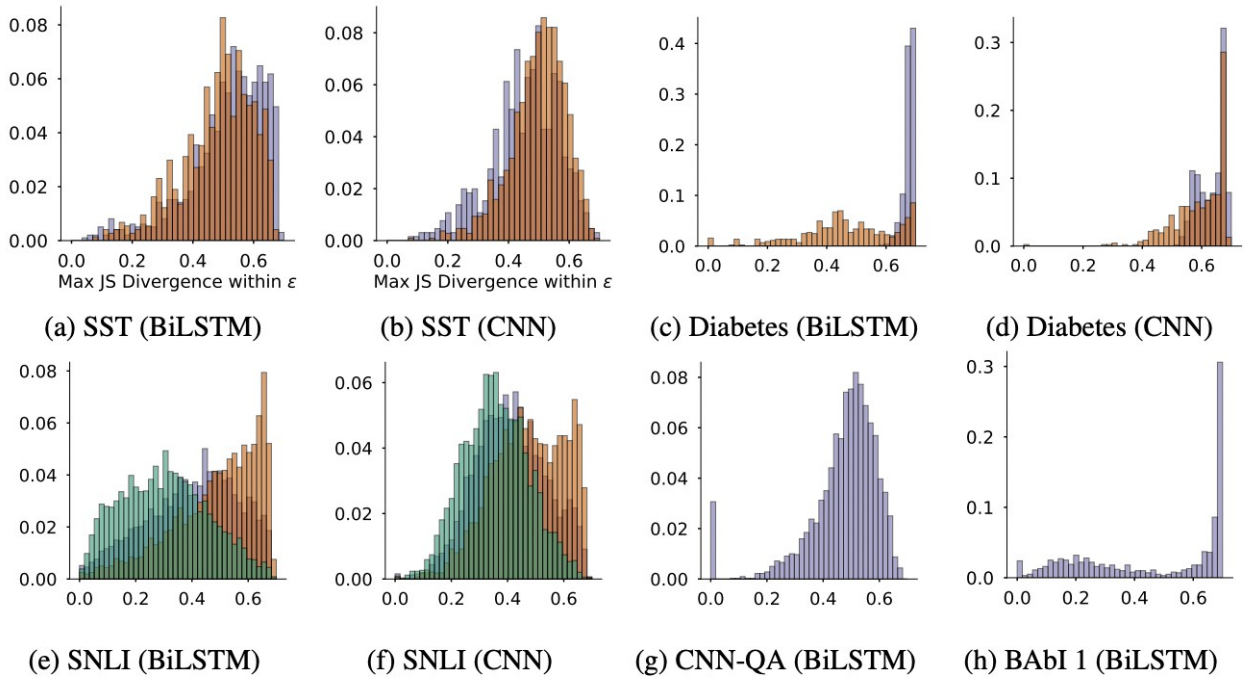


Figure 7: Histogram of **maximum adversarial JS Divergence** (**ϵ-max JSD**) between original and adversarial attentions over all instances. In all cases shown, $|\hat{y}^{adv} - \hat{y}| < \epsilon$. Encoders are specified in parantheses.

Adversarial Attention - Results

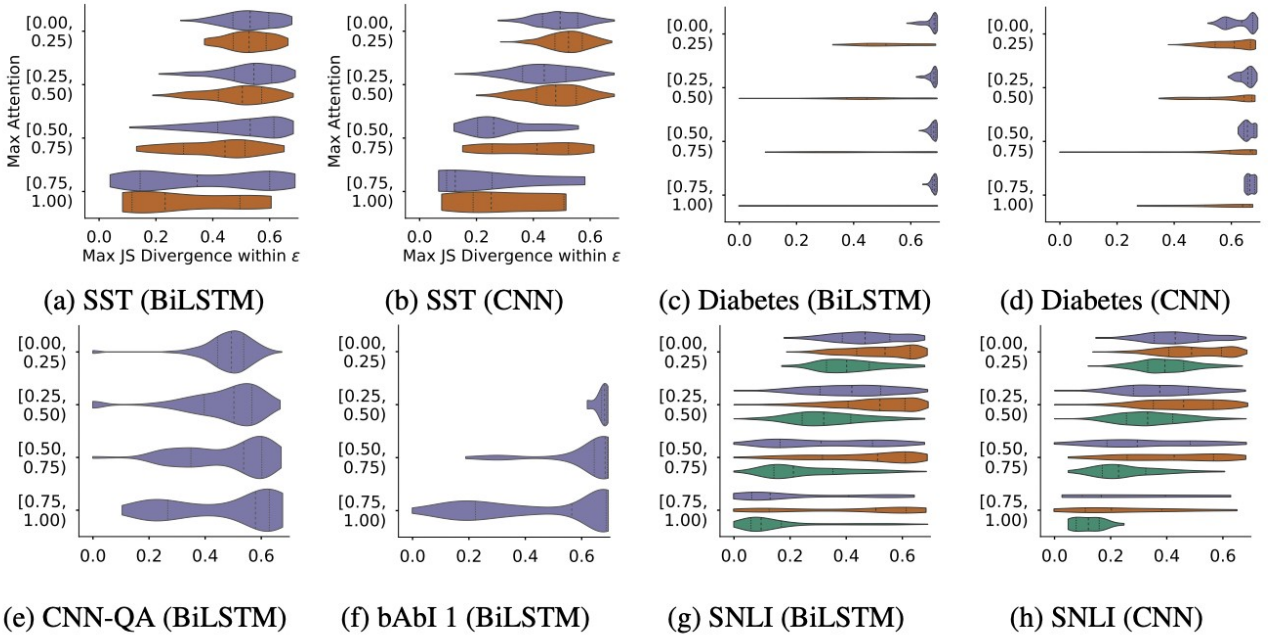


Figure 8: Densities of **maximum JS divergences** (**ϵ-max JSD**) (x-axis) as a function of the **max attention** (y-axis) in each instance for obtained between original and adversarial attention weights.

Counterfactual Attention - Limitations

- Adversarial weights may themselves be unlikely under model parameters
- There are instances where a single explanation out of multiple plausible explanations is sufficient.

Overall Limitations

- Alternate attention mechanisms are not explored
- Limited evaluations to classification, QA, and NPI tasks. seq2seq tasks are not considered.

Critique and Discussion

- Wiegreffe, S. & Pinter, Y. (2019). Attention is not not Explanation. In Proceedings of the 2019 Conference on EMNLP-IJCNLP (pp. 11-20).
  - Attention distribution is not a primitive.
  - Is attention necessary for prediction ?
  - Is the variance in attention distributions unusual ? i.e; How adversarial are the adversaries ?
- Do you agree with the authors' conditions of a "faithful explanation"?
- Do you think attention constitutes a "sufficient" explanation ?
- What other conditions should we consider as a good explanation?
- How can experts be brought into evaluate conditions of good explanations?