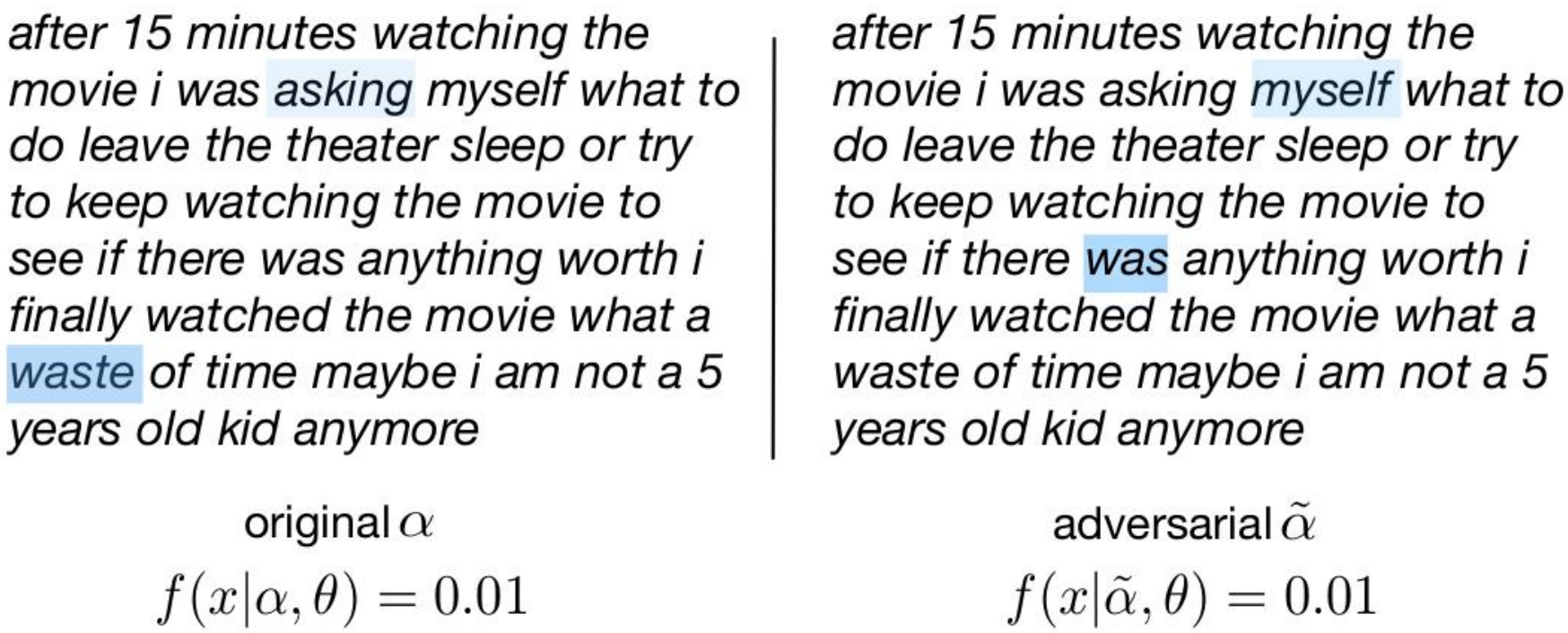
Attention is not Explanation

NAACL'19

Sarthak Jain, Byron C. Wallace
Northeastern University

Question

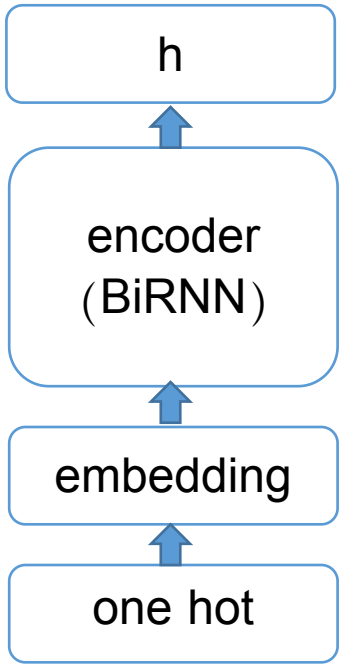
- Is the attention mechanism really get the semantic attention？

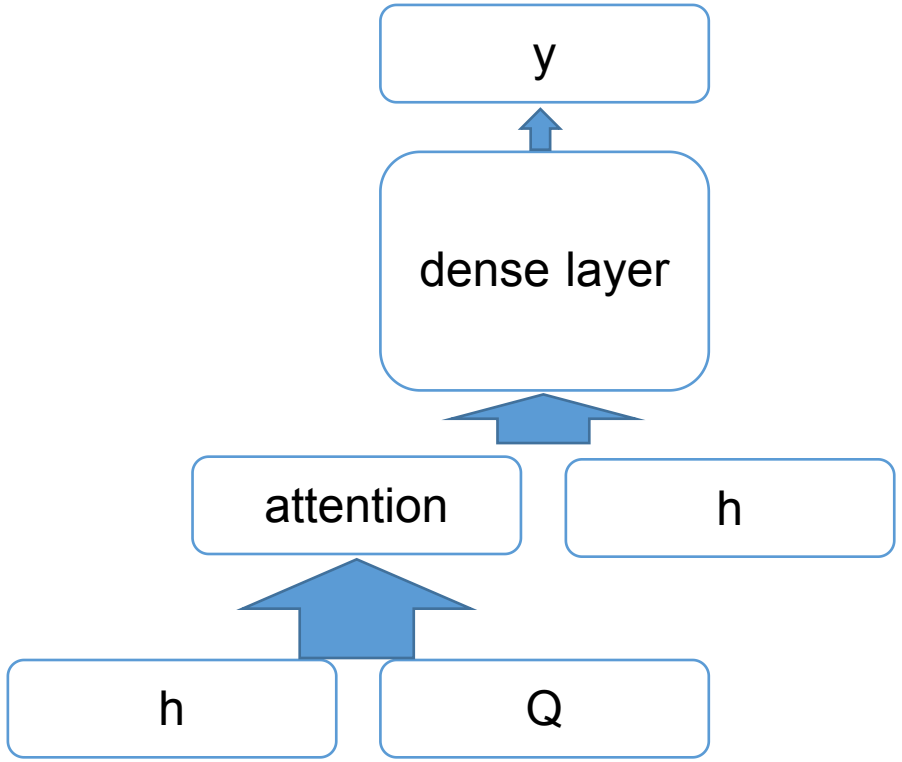


Is the attention provide transparency？

- Do attention weights correlate with measures of feature importance？(e.g., gradient-based measures);
- Would alternative attention weights necessarily yield different predictions？

Experiment Model





Dataset

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 / 0.48 / 0.62
SNLI	20982	14	182764 / 183187 / 183416	3219 / 3237 / 3368	0.78

Correlation with Feature Importance

- Gradient based measure
- Leave one feature out

Algorithm 1 Feature Importance Computations

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

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

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

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

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

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

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

Result for Correlation

- Gradients

Orange=|Positive, Purple=|Negative

O,P,G=|Neutral, Contradiction, Entailment

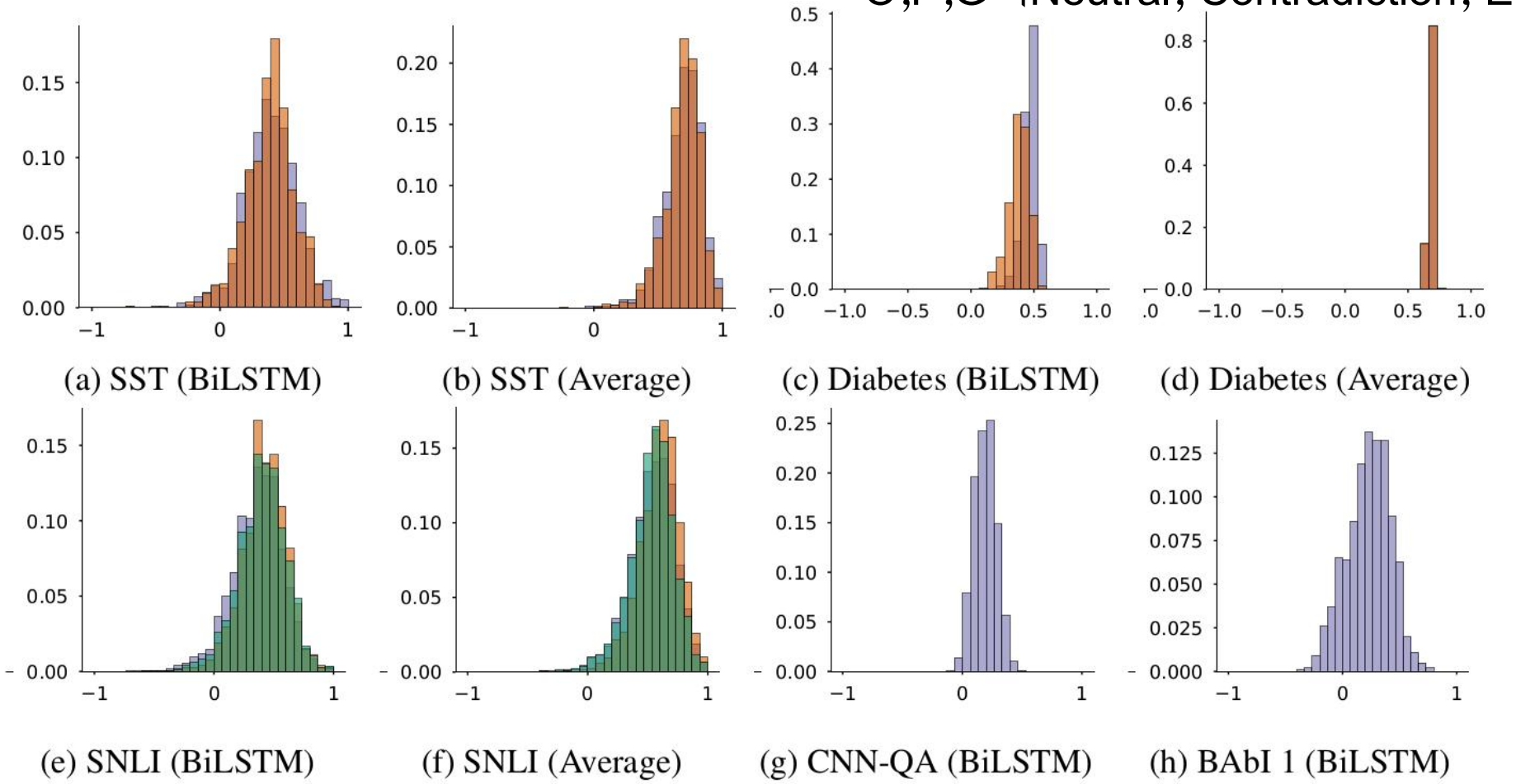


Figure 2: Histogram of **Kendall** τ between attention and gradients. Encoder variants are denoted parenthetically; colors indicate predicted classes. Exhaustive results are available for perusal online.

Statistically Significant

		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	0.40 ± 0.21	0.59	0.69 ± 0.15	0.93	0.34 ± 0.20	0.47
	1	0.38 ± 0.19	0.58	0.69 ± 0.14	0.94	0.33 ± 0.19	0.47
IMDB	0	0.37 ± 0.07	1.00	0.65 ± 0.05	1.00	0.30 ± 0.07	0.99
	1	0.37 ± 0.08	0.99	0.66 ± 0.05	1.00	0.31 ± 0.07	0.98
ADR Tweets	0	0.45 ± 0.17	0.74	0.71 ± 0.13	0.97	0.29 ± 0.19	0.44
	1	0.45 ± 0.16	0.77	0.71 ± 0.13	0.97	0.40 ± 0.17	0.69
20News	0	0.08 ± 0.15	0.31	0.65 ± 0.09	0.99	0.05 ± 0.15	0.28
	1	0.13 ± 0.16	0.48	0.66 ± 0.09	1.00	0.14 ± 0.14	0.51
AG News	0	0.42 ± 0.11	0.93	0.77 ± 0.08	1.00	0.35 ± 0.13	0.80
	1	0.35 ± 0.13	0.81	0.75 ± 0.07	1.00	0.32 ± 0.13	0.73
Diabetes	0	0.47 ± 0.06	1.00	0.68 ± 0.02	1.00	0.44 ± 0.07	1.00
	1	0.38 ± 0.08	1.00	0.68 ± 0.02	1.00	0.38 ± 0.08	1.00
Anemia	0	0.42 ± 0.05	1.00	0.81 ± 0.01	1.00	0.42 ± 0.05	1.00
	1	0.43 ± 0.06	1.00	0.81 ± 0.01	1.00	0.44 ± 0.06	1.00
CNN	Overall	0.20 ± 0.06	0.99	0.48 ± 0.11	1.00	0.16 ± 0.07	0.95
bAbI 1	Overall	0.23 ± 0.19	0.46	0.66 ± 0.17	0.97	0.23 ± 0.18	0.45
bAbI 2	Overall	0.17 ± 0.12	0.57	0.84 ± 0.09	1.00	0.11 ± 0.13	0.40
bAbI 3	Overall	0.30 ± 0.11	0.93	0.76 ± 0.12	1.00	0.31 ± 0.11	0.94
SNLI	0	0.36 ± 0.22	0.46	0.54 ± 0.20	0.76	0.44 ± 0.18	0.60
	1	0.42 ± 0.19	0.57	0.59 ± 0.18	0.84	0.43 ± 0.17	0.59
	2	0.40 ± 0.20	0.52	0.53 ± 0.19	0.75	0.44 ± 0.17	0.61

Table 2: Mean and std. dev. of correlations between gradient/leave-one-out importance measures and attention weights. *Sig. Frac.* columns report the fraction of instances for which this correlation is statistically significant;

Random Attention Weights

Algorithm 2 Permuting attention weights

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

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

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

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

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

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

end for

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

Result for Random Permutation

Orange=|Positive, Purple=|Negative O,P,G=|Neutral, Contradiction, Entailment

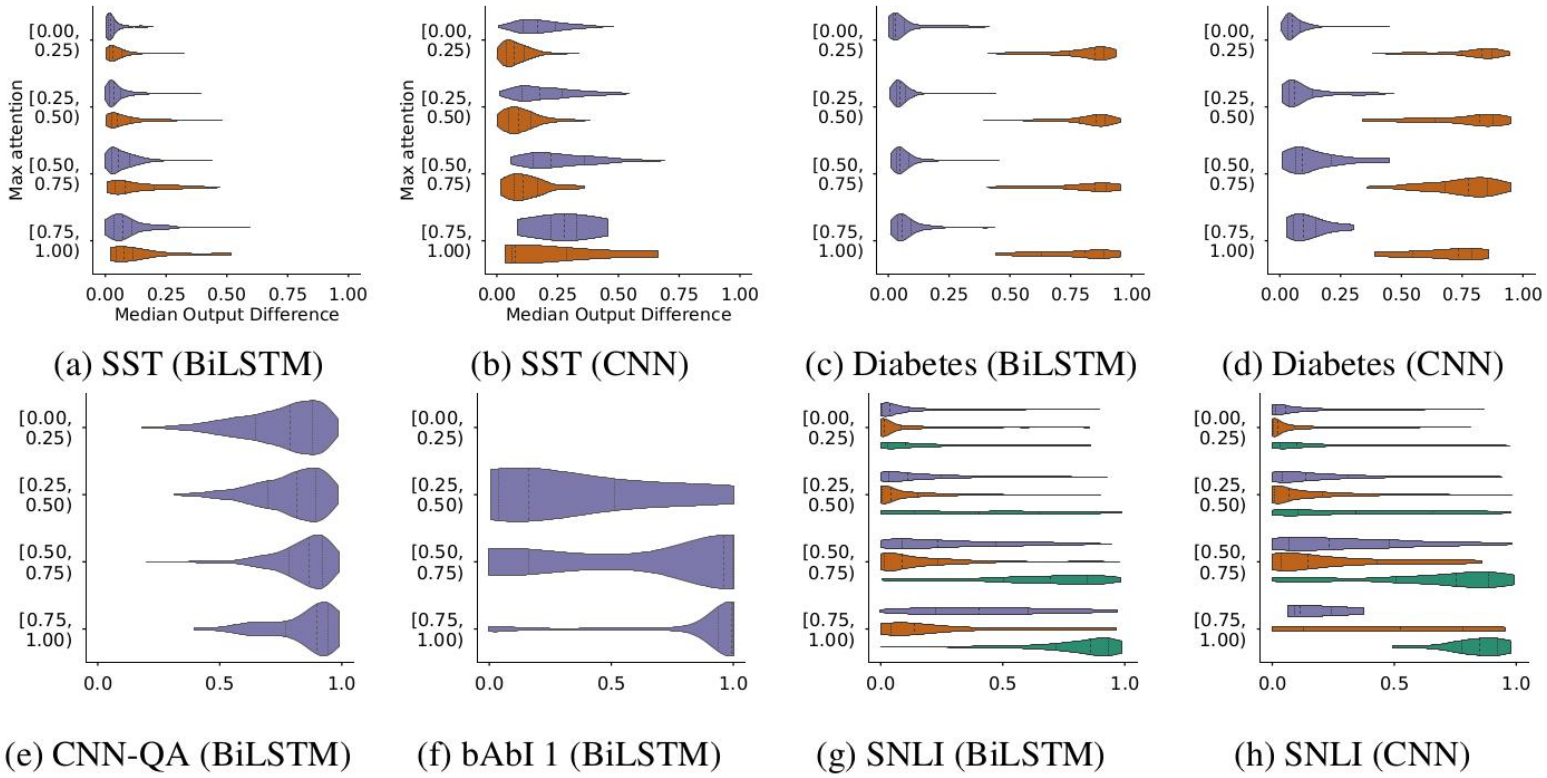


Figure 6: **Median change in output** $\Delta\hat{y}^{med}$ (x-axis) densities in relation to the **max attention** ($\max \hat{\alpha}$) (y-axis) obtained by randomly permuting instance attention weights. Encoders denoted parenthetically. Plots for all corpora and using all encoders are available online.

Adversarial Attention

- Optimize a relaxed version with Adam SGD

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

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

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

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

Result for Adversarial Attention

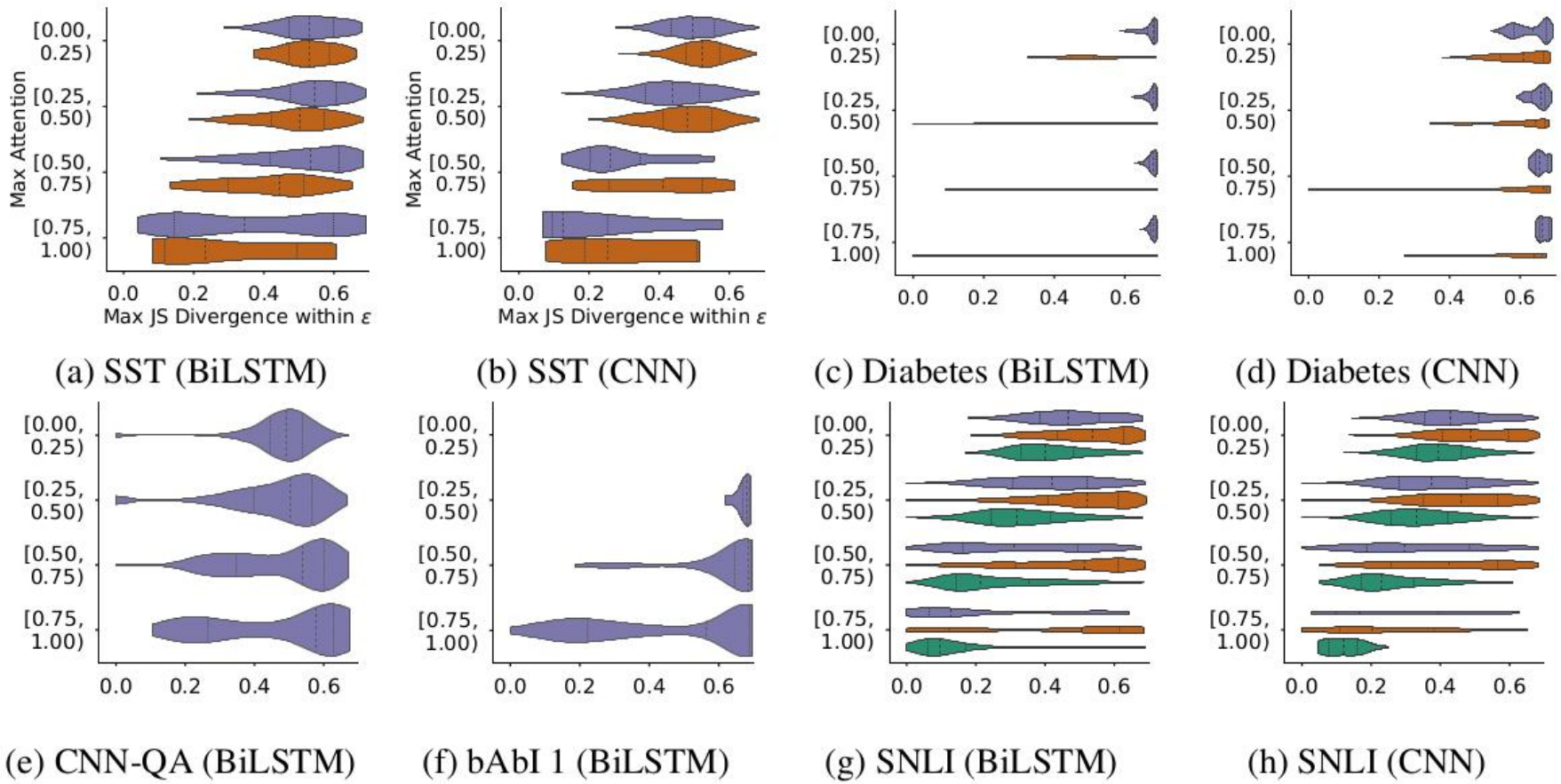


Figure 8: Densities of **maximum JS divergences** (**ϵ-max JSD**) (x-axis) as a function of the **max attention** (y-axis) in each instance for obtained between original and adversarial attention weights.

Conclusion

- correlation between feature importance measures and learned attention weights is weak
- counterfactual attentions often have no effect on model output
- limitations
  - only consider a handful of attention variants
  - only evaluate tasks with unstructured output spaces (no seq2seq)

Adversarial Heatmaps Example

SST

Original: reggio falls victim to relying on the very digital technology that he fervently scorns creating a meandering inarticulate and ultimately disappointing film

Adversarial: reggio falls victim to relying on the very digital technology that he fervently scorns creating a meandering inarticulate and ultimately disappointing film $\Delta\hat{y}$: *0.005*

IMDB

Original: fantastic movie one of the best film noir movies ever made bad guys bad girls a jewel heist a twisted morality a kidnapping everything is here jean has a face that would make bogart proud and the rest of the cast is is full of character actors who seem to to know they're onto something good get some popcorn and have a great time

Adversarial: fantastic movie one of the best film noir movies ever made bad guys bad girls a jewel heist a twisted morality a kidnapping everything is here jean has a face that would make bogart proud and the rest of the cast is is full of character actors who seem to to know they're onto something good get some popcorn and have a great time $\Delta\hat{y}$: *0.004*

Adversarial Heatmaps Example

AG News

Original:general motors and daimlerchrysler say they # qqq teaming up to develop hybrid technology for use in their vehicles . the two giant automakers say they have signed a memorandum of understanding

Adversarial:general motors and daimlerchrysler say they # qqq teaming up to develop hybrid technology for use in their vehicles . the two giant automakers say they have signed a memorandum of understanding . $\Delta\hat{y}$: *0.006*

SNLI

Hypothesis:a man is running on foot

Original Premise Attention:a man in a gray shirt and blue shorts is standing outside of an old fashioned ice cream shop named sara 's old fashioned ice cream , holding his bike up , with a wood like table , chairs , benches in front of him .

Adversarial Premise Attention:a man in a gray shirt and blue shorts is standing outside of an old fashioned ice cream shop named sara 's old fashioned ice cream , holding his bike up , with a wood like table , chairs , benches in front of him . $\Delta\hat{y}$: *0.002*