Attention is not Explanation

arXiv: https://arxiv.org/pdf/1902.10186.pdf

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

  → Only weakly and inconsistently
- Would alternative attention weights and hence distinct heatmaps/explanations necessarily yield different predictions?

  → No; it is very often possible to construct adversarial attention distributions that yield effectively equivalent predictions as when using the originally induced attention weights

Notations & Measures:

- inputs $\mathbf{x} \in \mathbb{R}^{T \times |V|}$

  → Embedding matrix $\mathbf{E}$
  - Token representations $\mathbf{x}_e \in \mathbb{R}^{T \times d}$

    → Encoder (Bi-RNN)
    - Hidden states: $\mathbf{h} = \mathbf{Enc}\left(\mathbf{x}_e\right) \in \mathbb{R}^{T \times m}$

      → Similarity function ϕ (**H**idden states, **Q**uery - question or hypothesis)
      - $\hat{\boldsymbol{\alpha}} = \operatorname{softmax}(\phi(\mathbf{h}, \mathbf{Q})) \in \mathbb{R}^T$

        → A dense layer taking the Hidden states & the attention weight
        - $\hat{y} = \sigma\left(\boldsymbol{\theta} \cdot h_\alpha\right) \in \mathbb{R}^{|\mathcal{Y}|}$ with $h_\alpha = \sum_{t=1}^T \hat{\alpha} t \cdot h t$

- Distance between two output distribution is measure using Total Variation Distance (TVD):

$$\operatorname{TVD}\left(\hat{y} 1, \hat{y} 2\right) = \frac{1}{2} \sum_{i=1}^{|\mathcal{Y}|} \left|\hat{y} 1 i - \hat{y} 2 i\right|$$

- Distance between two attention distributions is measured using Jensen-Shannon Divergence (JSD):

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

Experiments

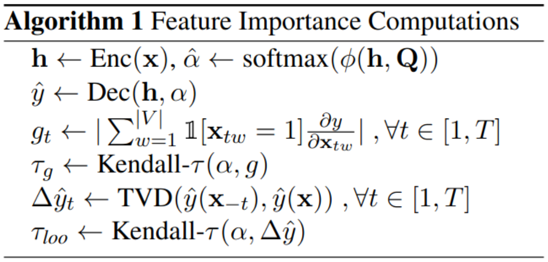
The objective is to answer "Do learned attention weights agree with alternative, natural measures of feature importance? And, Had we attended to different features, would the prediction have been different?"

Correlation Between Attention and Feature Importance Measures

We measure correlations between attention and:

(1) gradient-based measures of feature importance (τ_g),

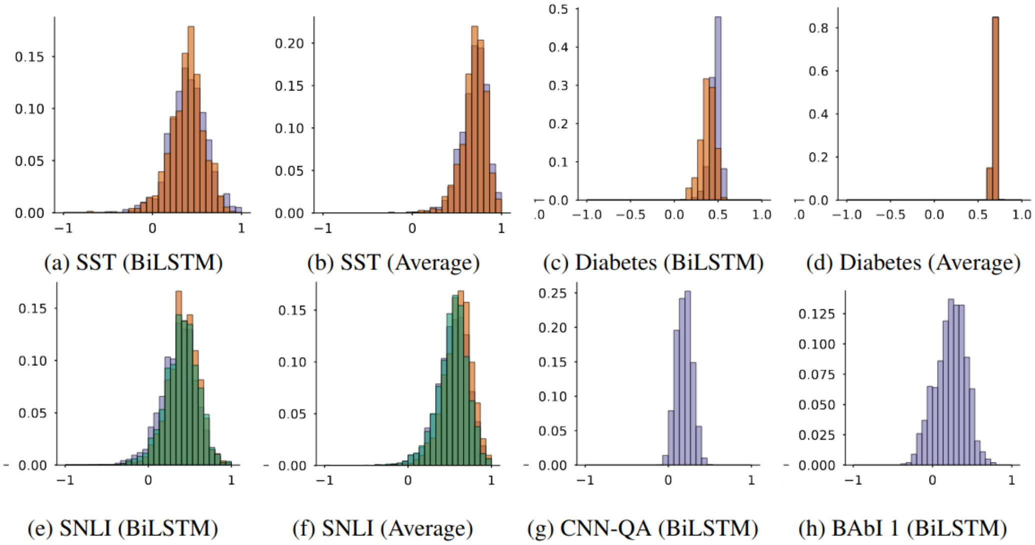
(2) differences in model output induced by leaving features out (τ_{loo}).



Algorithm 1 Feature Importance Computations

```
$\mathbf{h} \leftarrow \mathrm{Enc}(\mathbf{x}),\ \hat{\alpha} \leftarrow \mathrm{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \mathrm{Dec}(\mathbf{h}, \alpha)$
$g_t \leftarrow |\sum_{w=1}^{|V|} \mathbb{1}[\mathbf{x}_{tw} = 1] \frac{\partial y}{\partial \mathbf{x}_{tw}}|\ , \forall t \in [1, T]$
$\tau_g \leftarrow \text{Kendall-}\tau(\alpha, g)$
$\Delta\hat{y}_t \leftarrow \mathrm{TVD}(\hat{y}(\mathbf{x}_{-t}), \hat{y}(\mathbf{x}))\ , \forall t \in [1, T]$
$\tau_{loo} \leftarrow \text{Kendall-}\tau(\alpha, \Delta\hat{y})$
```

Note: for gradient computation, the computation graph is cut-off at the attention module so that gradient does not flow through this layer and contribute to the gradient feature importance score, this is because we want to measure: how much does the output change as we perturb particular inputs (words) by a small amount while paying the same amount of attention to the said word as originally estimated and shown in the heatmap?



(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)

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.

→ In general, observed correlations are modest (a value of 0 indicates no correspondence, while 1 implies perfect concordance) for the BiLSTM model. The centrality of observed densities hovers around or below 0.5 in most of the corpora considered.

How does LOO and gradients correlate with one another?

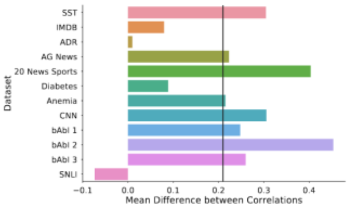


Figure 3: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. LOO scores using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by >0.2 τ_{loo}.

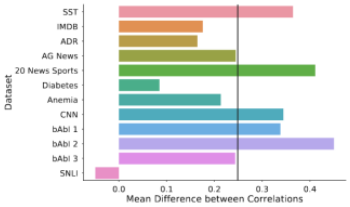


Figure 4: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. Gradients using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by $\sim$0.25 τ_g.

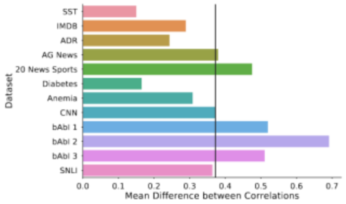


Figure 5: Difference in mean correlation of attention weights vs. LOO importance measures for (i) Average (feed-forward projection) and (ii) BiLSTM Encoders with Tanh attention. Average correlation (vertical bar) is on average $\sim$0.375 points higher for the simple feedforward encoder, indicating greater correspondence with the LOO measure.

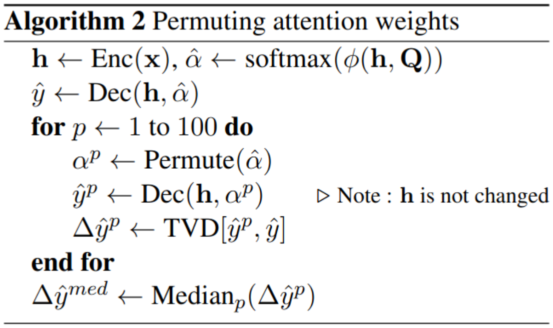
→ we find that they exhibit, in general, a considerably higher correlation with one another (on average) than LOO / Gradient does with attention scores

→ The results suggest that, in general, attention weights do not strongly or consistently agree with standard feature importance scores

Counterfactual Attention Weights

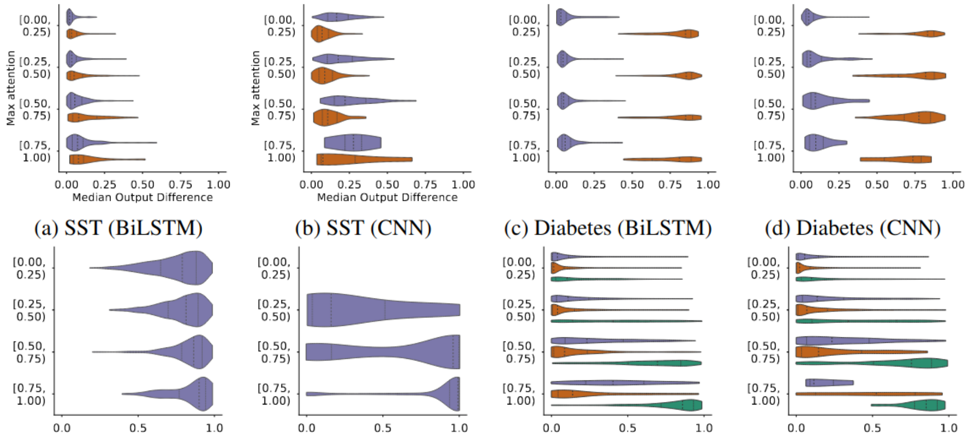
The idea is to investigate whether the prediction would have been different, had the model emphasized (attended to) different input features. This consists of considering an alternative attention distribution and the output distribution induced by it and compare it to the original output distribution, telling us the degree to which a particular (attention) heat map uniquely induces an output.

1- Attention Permutation: simply scramble the original attention weights, re-assigning each value to an arbitrary, randomly sampled index (input feature).



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



(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)

Figure 6: **Median change in output $\boldsymbol{\Delta\hat{y}^{med}}$** (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.

→ We observe that there exist many points with small Δy despite large magnitude attention weights. These are cases in which the attention weights might suggest explaining an output by a small set of features (this is how one might reasonably read a heatmap depicting the attention weights), but were scrambling the attention makes little difference to the prediction.

2- Adversarial Attention: The intuition is to explicitly seek out attention weights that differ as much as possible from the observed attention distribution and yet leave the prediction effectively unchanged

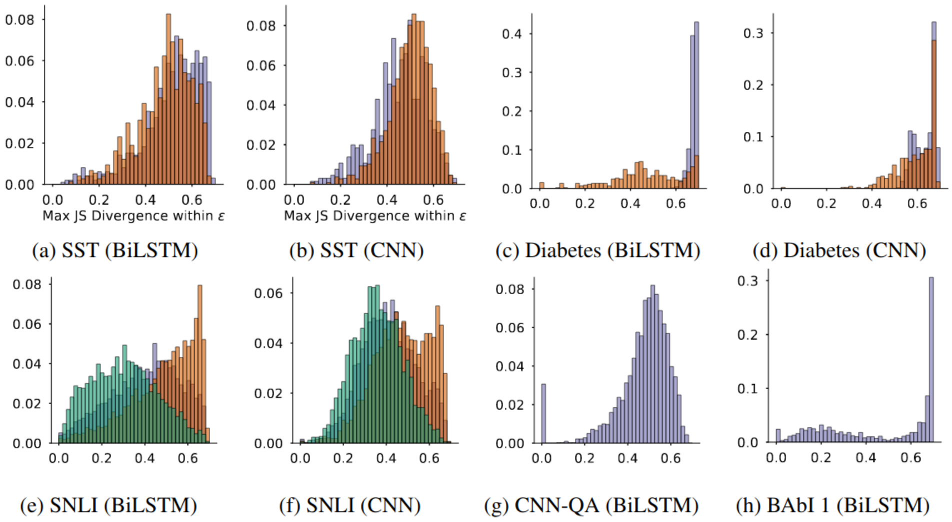
$$\begin{aligned} &\underset{\alpha^{(1)},\ldots,\alpha^{(k)}}{\text{maximize}} \quad f(\{\alpha^{(i)}\}_{i=1}^{k}) \\ &\text{subject to} \quad \forall i \ \text{TVD}[\hat{y}(\mathbf{x},\alpha^{(i)}),\hat{y}(\mathbf{x},\hat{\alpha})] \leq \epsilon \end{aligned} \tag{1}$$

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

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

Algorithm 3 Finding adversarial attention weights

```
$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x})$, $\hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h},\mathbf{Q}))$
$\hat{y} \leftarrow \text{Dec}(\mathbf{h},\hat{\alpha})$
$\alpha^{(1)},\ldots,\alpha^{(k)} \leftarrow$ 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
$\epsilon$-max JSD $\leftarrow \max_i \mathbb{1}[\Delta\hat{y}^{(i)} \leq \epsilon]\Delta\alpha^{(i)}$
```



(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)

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.

→ it is often the case that quite different attention distributions over inputs would yield essentially the same (within) output