Attention is not Explanation

Sarthak Jain, Byron C. Wallace - Northeastern University

21 Feb 2020

Presenter: Sanchit Sinha

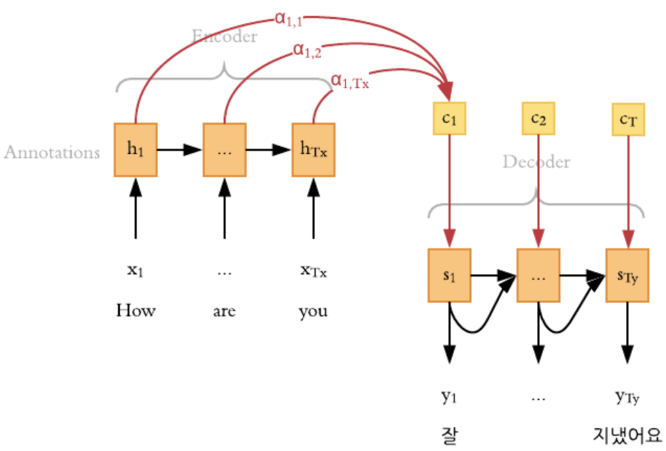
https://qdata.github.io/deep2Read/

Motivation

- **Attention mechanisms** are being used to demonstrate transparency in standard NLP downstream tasks - text classification, question answering and natural language inference
- Is attention **actually explaining** the outputs of models trained for such tasks?
- If yes, perform extensive experiments to assess the degree to which attention weights provide “meaningful explanations" for predictions
- Similar in essence to the sanity check paper - experiment idea and design is similar

Background

- **Attention methods** have been shown to improve upon the performance of standard encoder-decoder architectures
- Intuitive figure demonstrating attention in machine translation:



- Global vs Local attention: Output of one “token” in the output is dependent on all the hidden units in a weighted fashion (Global) or only on a few of the hidden units (Local)
- **Why Attention?** To capture a much more holistic dependence on the output with respect to hidden states

Background

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

- Jensen Shannon Divergence: $\mathrm{JSD}(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$ $M = \frac{1}{2}(P + Q)$

- For Correlation measurement : Kendal Tau
- Encoder Model:
    - Average - simple
    - BiLSTM - recurrent

Related Work

- Neural Machine Translation by Jointly Learning to Align and Translate - Bahdanau et al., 2014 (Attention Paper)
- A causal framework for explaining the predictions of black-box sequence-to-sequence models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing -David Alvarez-Melis and Tommi Jaakkola. 2017.
- An interpretable predictive model for healthcare using reverse time attention mechanism, Advances in Neural Information Processing Systems - Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart.

Claim / Target Task

- Comparison with other techniques:
  - **Correlation Between Attention and Feature Importance Measures -** does the attention weights have any correlation with the gradient-based methods of interpretability
- Modification of attention weights:
  - **Attention Permutation**- Permuting the weights of the attention on hidden states and checking if it makes a difference
  - **Adversarial Attention -** Adversarially computing new attention weights such that model predictions don't change a lot but attention weights change a lot.
- To perform these experiments over a variety of datasets on multiple tasks.
- https://successar.github.io/AttentionExplanation/docs/

Data Summary

Datasets used can be divided on the basis of the task:

- Binary text classification
  - Stanford Sentiment Treebank (SST)
  - IMDB Large Movie Reviews Corpus
  - Twitter Adverse Drug Reaction
  - 20 Newsgroups (Hockey vs Baseball).
  - AG News Corpus (Business vs World)
  - MIMIC ICD9 (Diabetes)
  - MIMIC ICD9 (Chronic vs Acute Anemia)
- Question Answering (QA)
  - CNN News Articles
  - bAbI
- Natural Language Inference
  - SNLI dataset

An Intuitive Figure Showing WHY Claim

$$x \xrightarrow[\text{Matrix}(E)]{\text{Embed}} x_e \xrightarrow{\text{Encoder}} h \xrightarrow{\phi} \phi(h, Q)$$

x: one-hot encoded $(T \times V)$, T: # of tokens, V: Vocab

x_e: $(T \times d)$

Encoder: Avg., BiLSTM

h: $(T \times m)$

$\phi(h, Q)$: R^T

Query (Q)

$$\hat{\alpha} = \text{softmax}(\phi(h, Q))$$

$$\Downarrow$$

$$\hat{y} = \sigma(\theta \cdot h_\alpha)$$

where

$$h_\alpha = \sum_{t=1}^{T} \hat{\alpha}_t h_t$$

Types of ϕ:

1. Additive $\Rightarrow$ $\phi(h, Q) = v^T \tanh(W_1 h + W_2 Q)$
2. Scaled Dot Product $\Rightarrow$ $\phi(h, Q) = \dfrac{hQ}{\sqrt{m}}$

Proposed Solution

- **Experiment-1** Correlation between Attention Weights and Gradient/LOO
- Calculating the correlation:
  - Tau_g -> corr. of gradients wrt attention weights
  - Tau_LOO -> corr. of leave one out wrt attention weights

Algorithm 1 Feature Importance Computations

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

Proposed Solution

- **Experiment-2**
  - Permuting Attention Weights

Algorithm 2 Permuting attention weights

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

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

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

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

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

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

end for

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

Proposed Solution

- **Experiment 2**
  - **Adversarial Attention -** “attention weights that differ as much as possible from the observed attention distribution and yet leave the prediction effectively unchanged.”
  - JS Divergence between any two categorical distributions irrespective of length) is bounded from above by 0.69.

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

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

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

Algorithm 3 Finding adversarial attention weights

$\mathbf{h} \leftarrow \mathrm{Enc}(\mathbf{x})$, $\hat{\alpha} \leftarrow \mathrm{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \mathrm{Dec}(\mathbf{h}, \hat{\alpha})$
$\alpha^{(1)}, \ldots, \alpha^{(k)} \leftarrow$ Optimize Eq 1
for $i \leftarrow 1$ to k **do**
  $\hat{y}^{(i)} \leftarrow \mathrm{Dec}(\mathbf{h}, \alpha^{(i)})$ ▷ **h** is not changed
  $\Delta\hat{y}^{(i)} \leftarrow \mathrm{TVD}[\hat{y}, \hat{y}^{(i)}]$
  $\Delta\alpha^{(i)} \leftarrow \mathrm{JSD}[\hat{\alpha}, \alpha^{(i)}]$
end for
ϵ-max JSD $\leftarrow \max_i \mathbb{1}[\Delta\hat{y}^{(i)} \leq \epsilon]\Delta\alpha^{(i)}$

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*

Experimental Results - Experiment 1

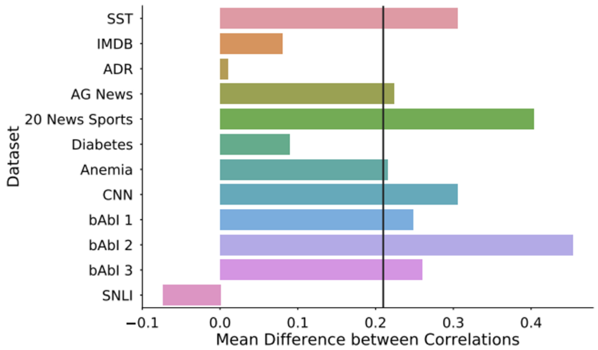


Figure 3: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. LOO scores using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by >0.2 τ_{loo}.

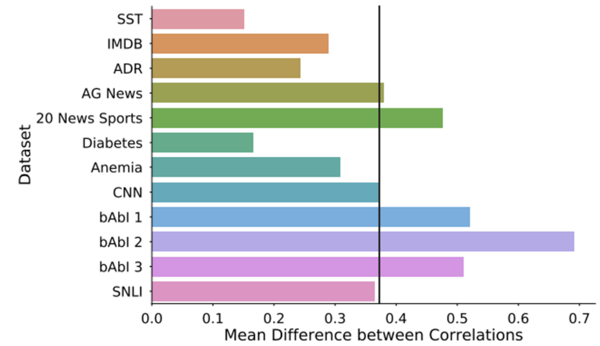


Figure 5: Difference in mean correlation of attention weights vs. LOO importance measures for (i) Average (feed-forward projection) and (ii) BiLSTM Encoders with Tanh attention. Average correlation (vertical bar) is on average ~0.375 points higher for the simple feedforward encoder, indicating greater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

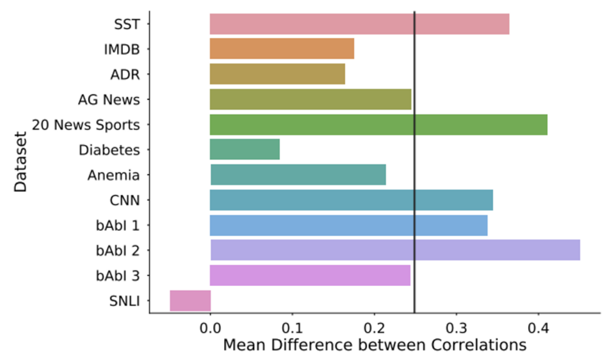
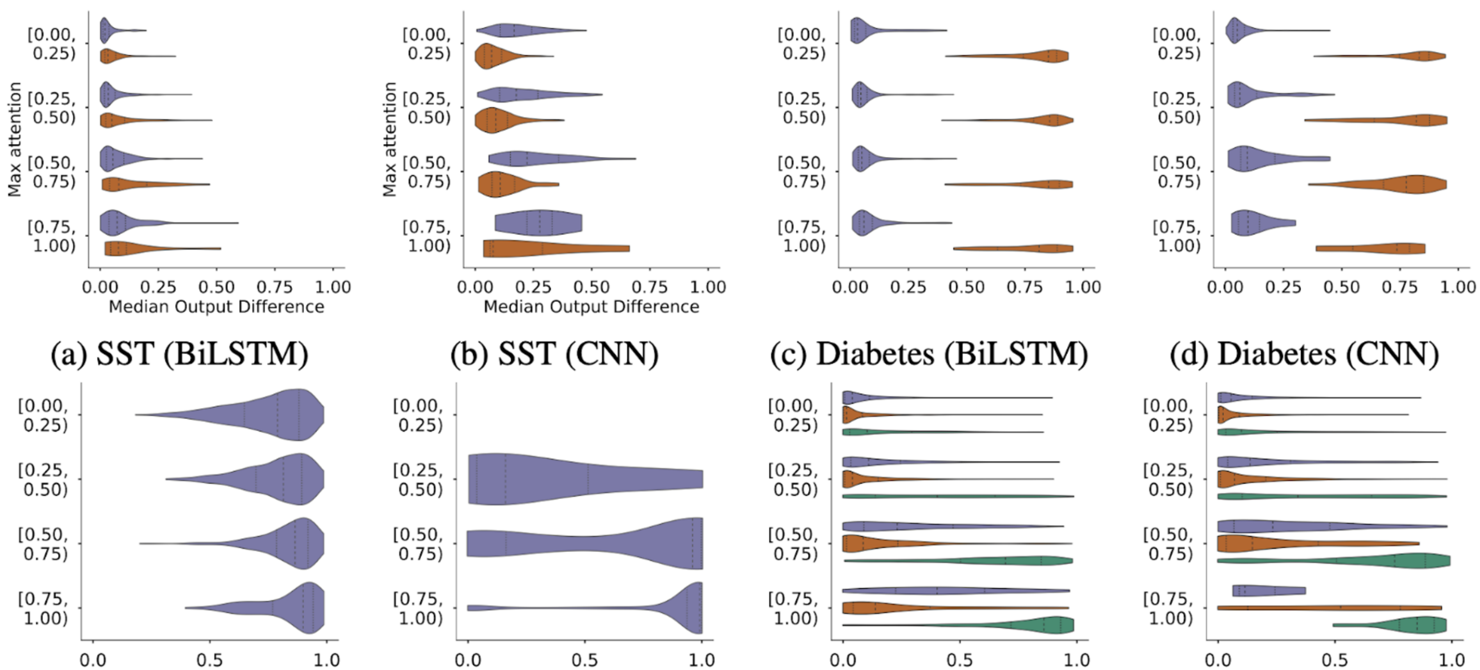


Figure 4: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. Gradients using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by ~0.25 τ_g.

Experimental Results - Experiment 2a



(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 $\Delta\hat{y}^{med}$** (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.

Experimental Results - Experiment 2b

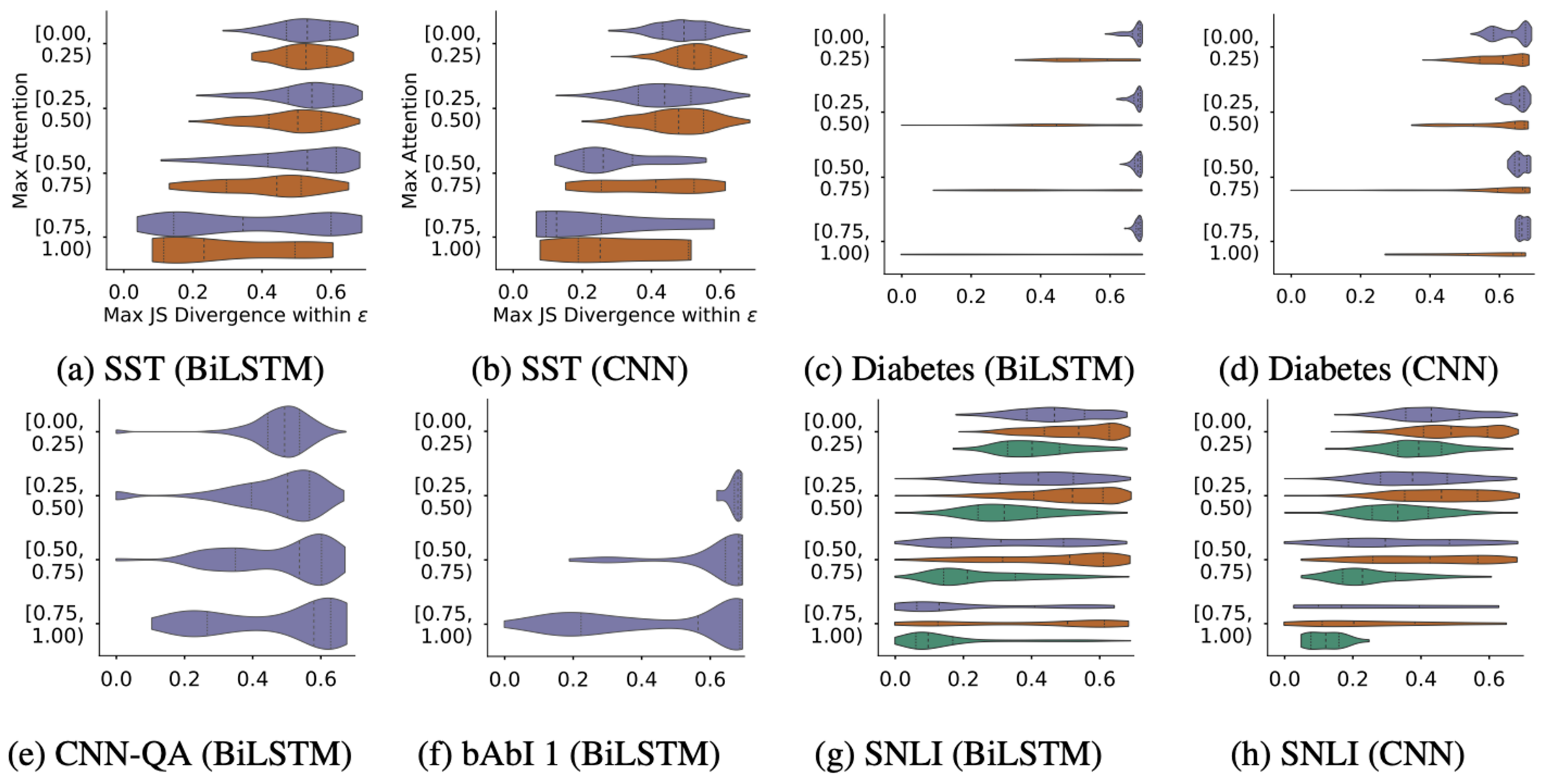


Figure 8: Densities of maximum JS divergences (-max JSD) (x-axis) as a function of the maxattention (y-axis) in each instance for obtained between original and adversarial attention weights.

Experimental Analysis

- **Experiment-1 :** Correlation study
  - Corr between LOO and Gradients is high
  - Corr between Gradients and attention and LOO and attention is on the lower side from expected
  - Corr of G/LOO vs attention for different encoders is different.
  - Simple encoders have high corr.(Average) and complex (BiLSTM) have low corr.
- **Experiment-2a:** Perturbing attention weights
  - The change in output by perturbing attention weights is much lower than expected
- **Experiment-2b:** Adversarial attention
  - “one can identify adversarial attention weights associated with high JSD for a significant number of examples. This means that it is often the case that quite different attention distributions over inputs would yield essentially the same output.

Conclusion and Future Work

- Showed that there is much more research required in studying attention
- Attention in itself is not enough to explain the models
- The failure of explainability of BiLSTM over average encoders is much more concerning due to the fact that still complex models are not very well understood