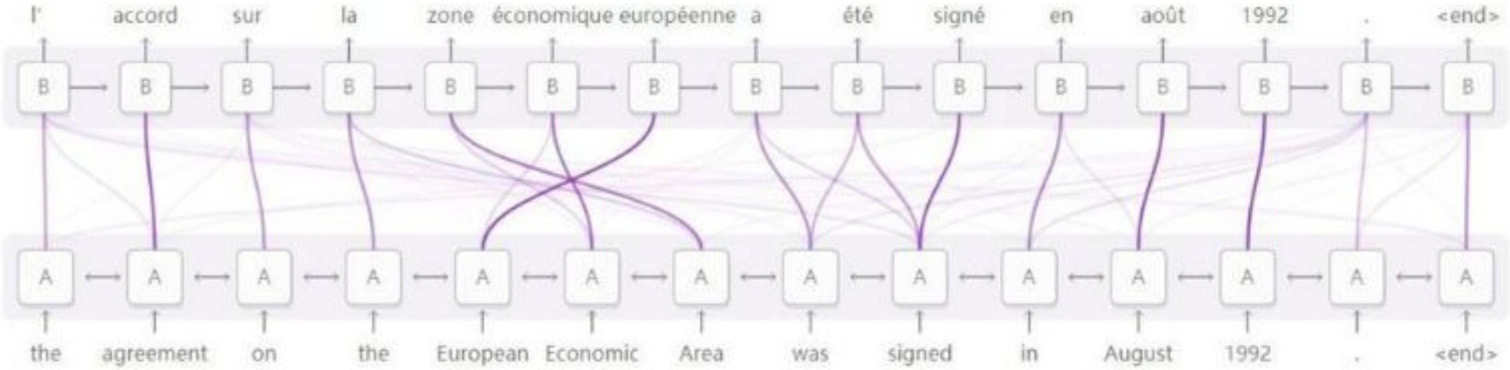
Attention is not Explanation

Nguyễn Đức Thắng

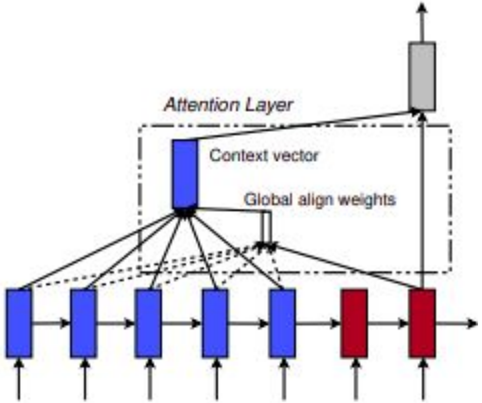
Background



l'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>
B
A
the
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>

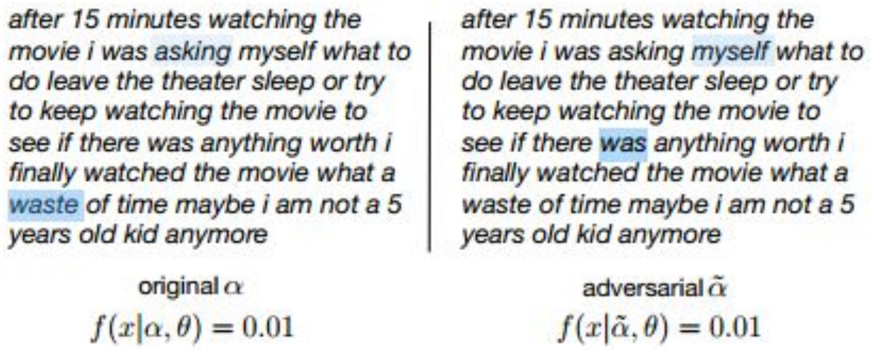
Background

- Given sequence h and query Q
- Calculate attention distribution: $\widehat{\alpha} = softmax(\phi(h, Q))$
  - Additive function: $\phi(h, Q) = v^T \tanh(W_1 h + W_2 Q)$
  - Scaled dot-product function: $\phi(h, Q) = \frac{hQ}{\sqrt{m}}$
- Get attention vector: $\alpha = h * \widehat{\alpha}$



Question

Is the attention mechanism really get semantic attention?



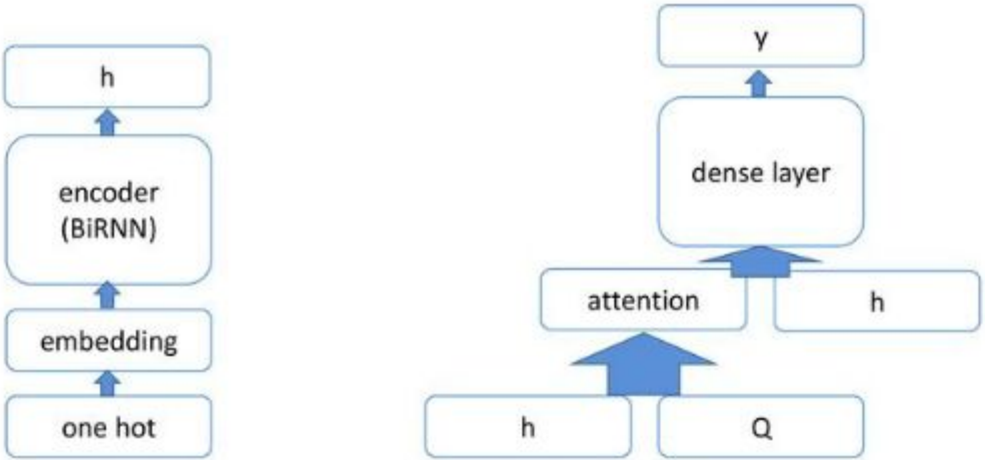
Is the attention provide transparency?

Premise: It’s easy to assume attention weights are measuring the effect of the input on the output. Is this quantifiable valid?

Hypothesis:

1. Attention weights should correlate with feature importance measures (e.g, gradient-based measures)
2. Alternative (or counterfactual) attention weight configurations ought to yeild corresponding changes in prediction (and if they do not then are equally plausible as explanations)

Experiment Model



Dataset

Dataset	$\lvert V \rvert$	*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

Experiment #1: Feature Importance Correlation

Research question: Do attention weights correlate with feature importance measure?

Measure:

- **Gradient-based methods**: The gradients of the model's output probabilities literally describe the model's decision boundary. How much do gradients change as a function of input, keeping attention weight fixed?
- **Leave-one-out**: A word's importance can be measured by the difference in model confidence before and after that word is removed from the input.

Experiment #1: Feature Importance Correlation

Algorithm 1 Feature Importance Computations

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

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

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

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

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

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

Result for Correlation

Orange=>Positive, Purple=>Negative

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

- Gradients

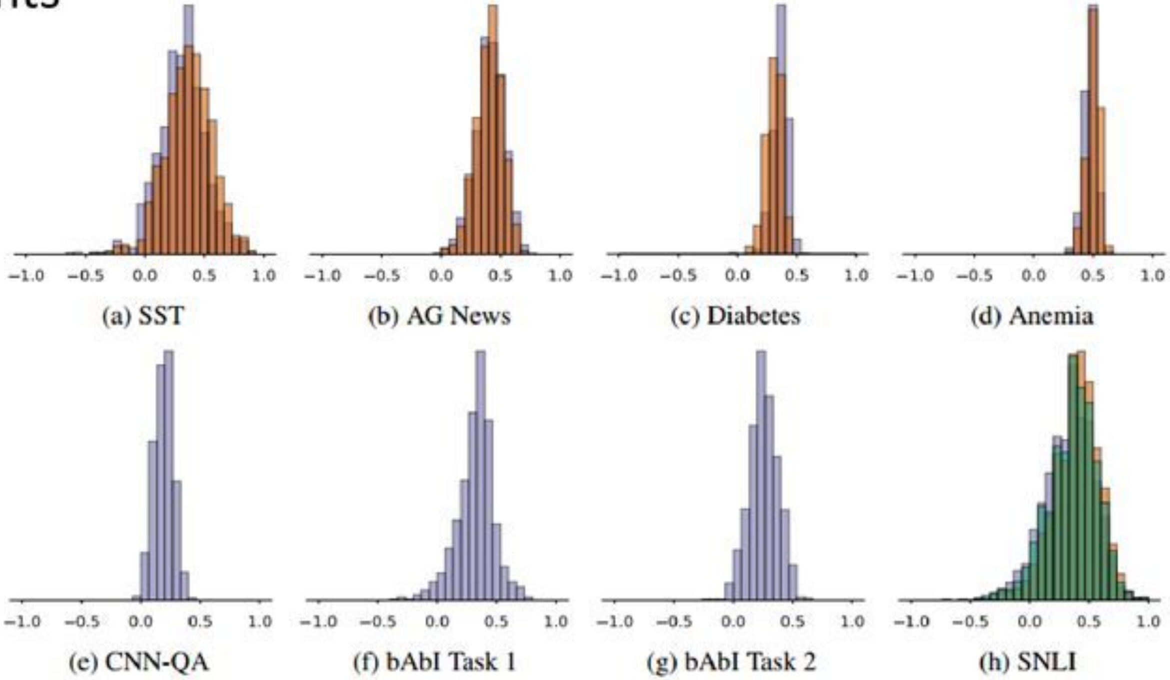


Figure 2: Histogram of **Kendall** τ between attention and gradients. Colors indicate the predicted classes.

Result for Correlation

- Leave One Out

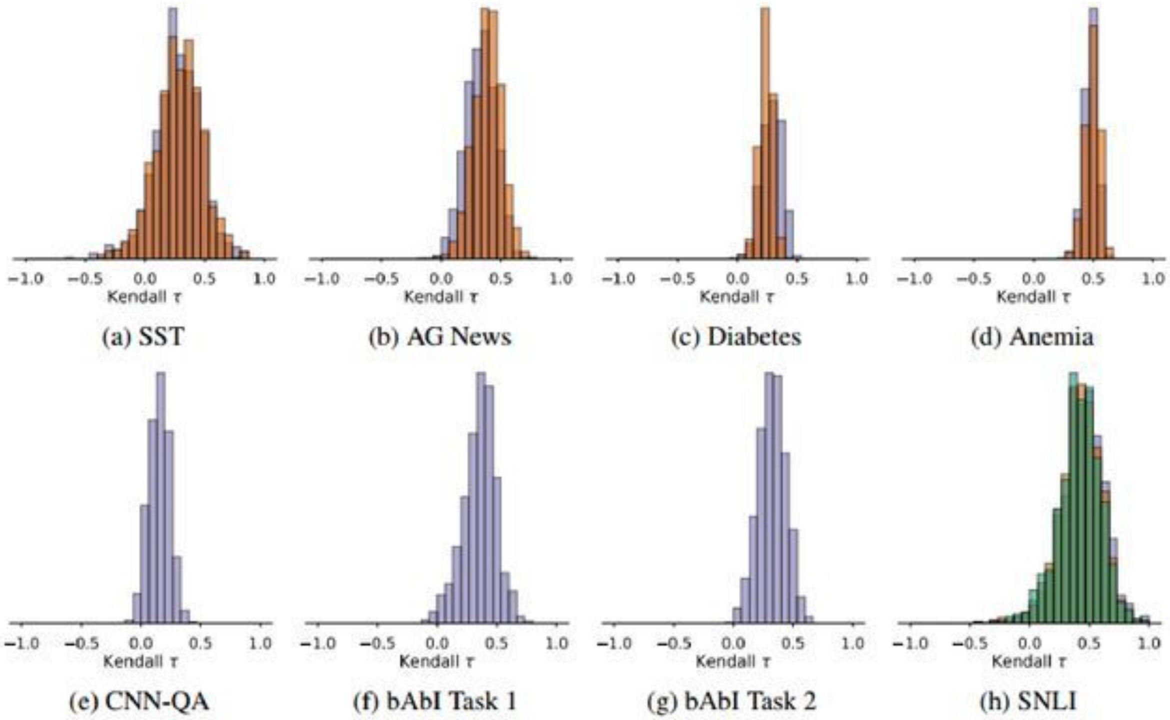


Figure 8: Kendall τ between attention and Leave-One-Out Importance Score w.r.t to input.

Experiment #2: Counterfactual Attention Weights

Attention Permutation: Randomly shuffling the elements of attention weights.

Adversarial Attention: Maximally perturbing attention weights while maintaining the same prediction.

Experiment #2: Counterfactual 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 : **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

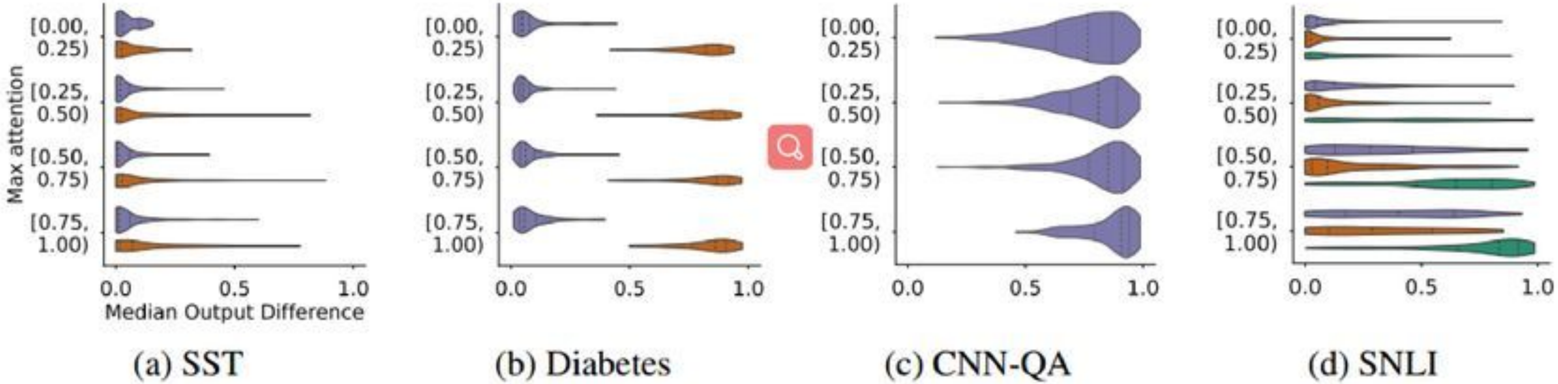


Figure 3: **Median change in output $\Delta\hat{y}^{med}$** (x-axis) densities in relation to the **max attention ($\max\hat{\alpha}$)** (y-axis) obtained by randomly permuting instance attention weights. Plots for all corpora are in the Appendix.

Experiment #2: Counterfactual Attention Weights

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

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

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

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

Result for Adversarial Attention

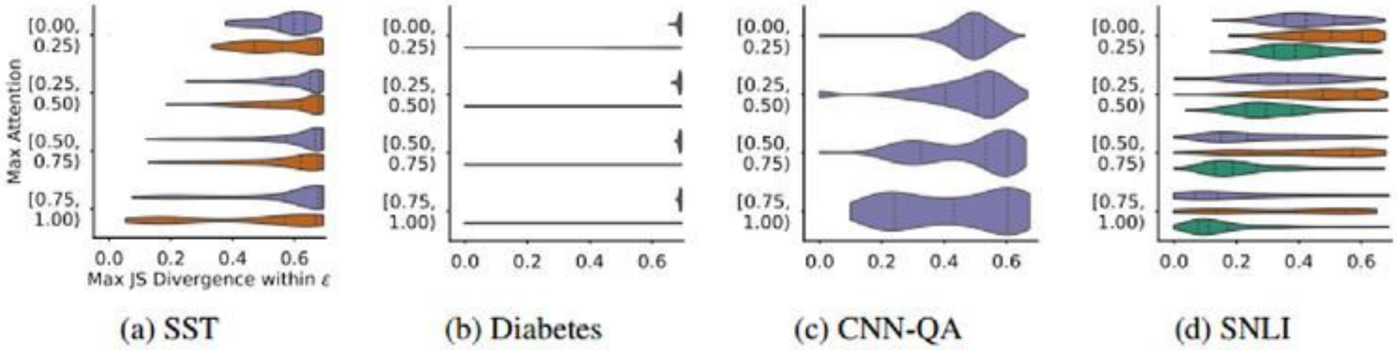


Figure 5: Densities of **maximum JS divergences (ϵ-max JSD)** (x-axis) as a function of the **max attention** (y-axis) in each instance for obtained between original and adversarial attention weights.

Conclusions

*Do learned attention weights agree with alternative, natural measures of feature importance? **Not significantly***

*And, had we attended to different features. would the prediction have been different? **Not at all***

Extensive research

The following papers are related extensions of this paper:

- *Attention is not not Explanation*
- *On Identifiability in Transformers*
- *Is Attention Interpretable?*

Thanks