3billion

Your one answer

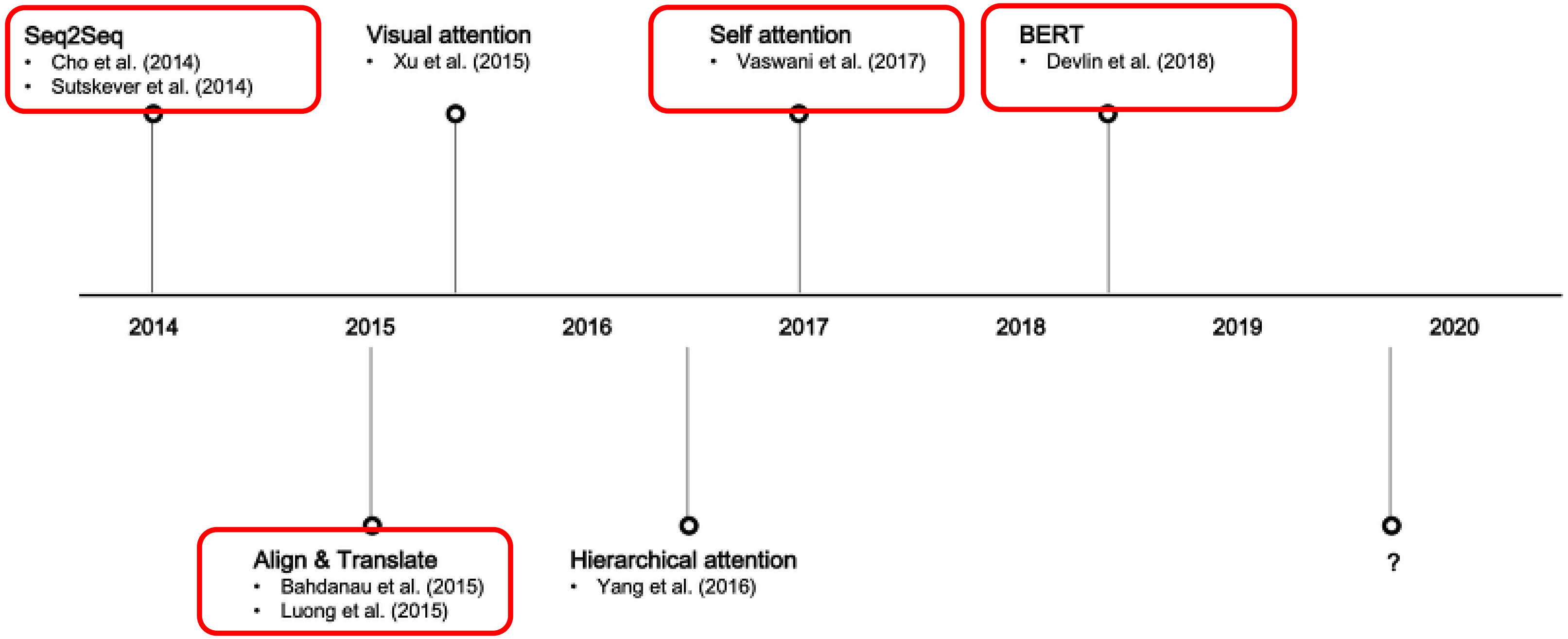
Attention is not Explanation

2023.04.04

Preliminary -1

- Attention mechanism(attention)
  - Attention ?
    - 자연어처리(machine translation)분야에서 처음으로 고안되어 많은 영역에서 사용됨
    - 긴 시퀀스에서의 입력시퀀스의 정보를 보존하기 위한 기법, Bahdaunau, Loung가 처음으로 고안
    - 현재는 Model transparency을 높이기 위해서 사용되고 있음

Figure 1. Key developments in attention



(2022)
ViT, SwinTransformer

개요

- Attention mechanism(attention)
    - Preliminary

Preliminary

3billion

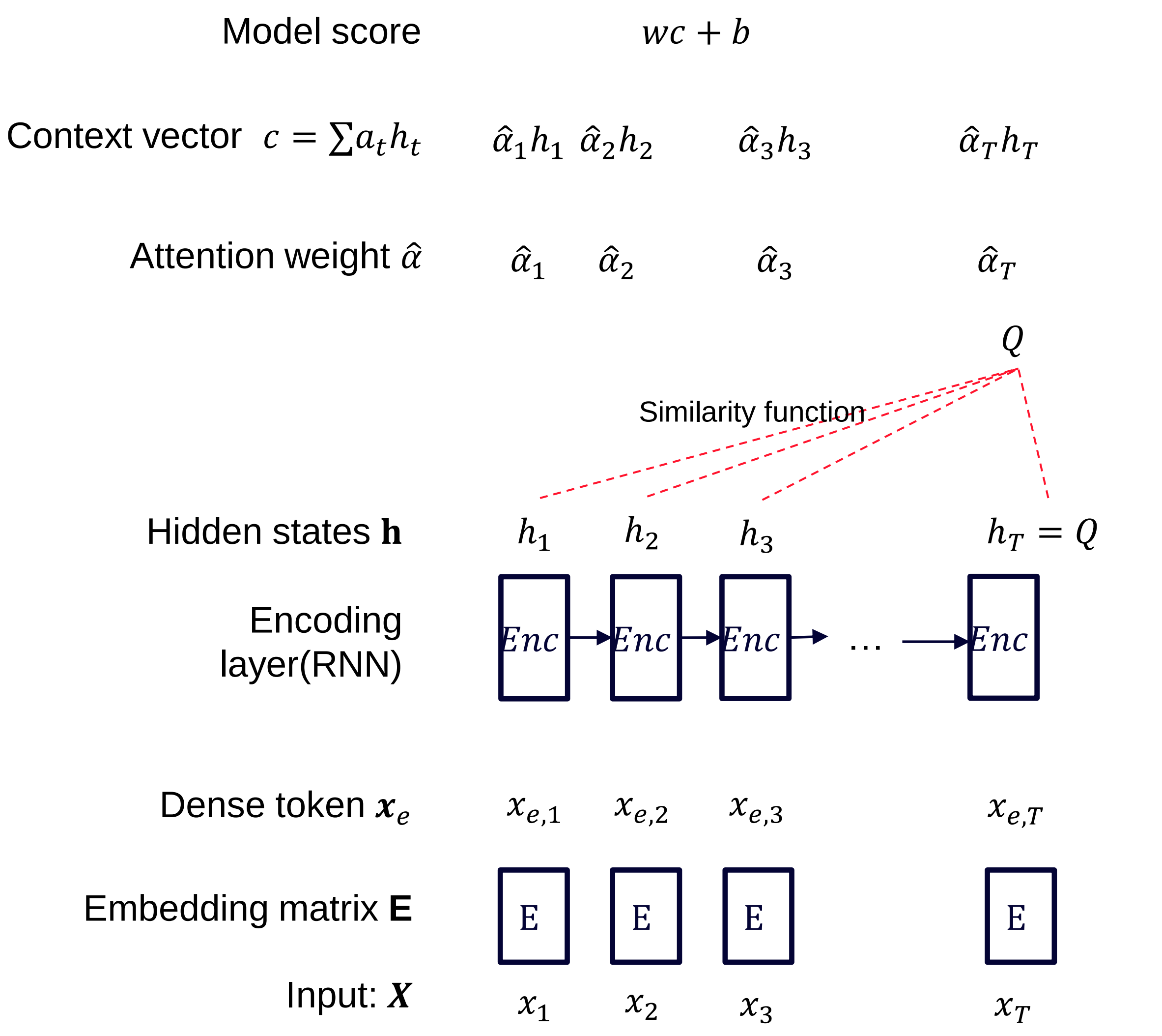
- Model input: $\boldsymbol{X} \in \mathbb{R}^{T\times|V|}$
  - T: timestamp
  - |V|: dimension of one-hot vector (one-hot encoded words)
- Embedding matrix **E** $\in \mathbb{R}^{|V|\times d}$
- Dense token : $\boldsymbol{x}_e \in \mathbb{R}^{T\times d}$
- Hidden states : $\mathbf{h} = Enc(x_e) \in \mathbb{R}^{T\times m}$

- Similarity function (alignment function) ϕ : $(\mathbf{h} \in \mathbb{R}^{T\times m}, Q \in \mathbb{R}^{m} \rightarrow \mathbb{R}^{T})$
  - Hidden representation: Q
  - Additive : $v^T \tanh(\mathrm{W}_1\mathbf{h} + \mathrm{W}_2 Q)$ $(\mathrm{v}, \mathrm{W}_1, \mathrm{W}_2$: trainable param) [2]
  - Scaled Dot product : $\frac{hQ}{\sqrt{m}}$ [3]
- Attention weight: $\hat{\boldsymbol{\alpha}} = \mathrm{softmax}\big(\phi(h, Q)\big) \in \mathbb{R}^{T}$.

- Dec
  - $\hat{y} = \sigma(\boldsymbol{\theta} \cdot h_\alpha)$
    - $h_\alpha = \sum_{t=1}^{T} \hat{a}_t \cdot h_t$
    - σ: activation function
    - $|\mathcal{Y}|$: label set size

[2]: Bahdanau et al, 2014
[3]: Vaswani et al 2017

Preliminary



[2] : Bahdanau et al, 2014
[3]: Vaswani et al 2017

Preliminary

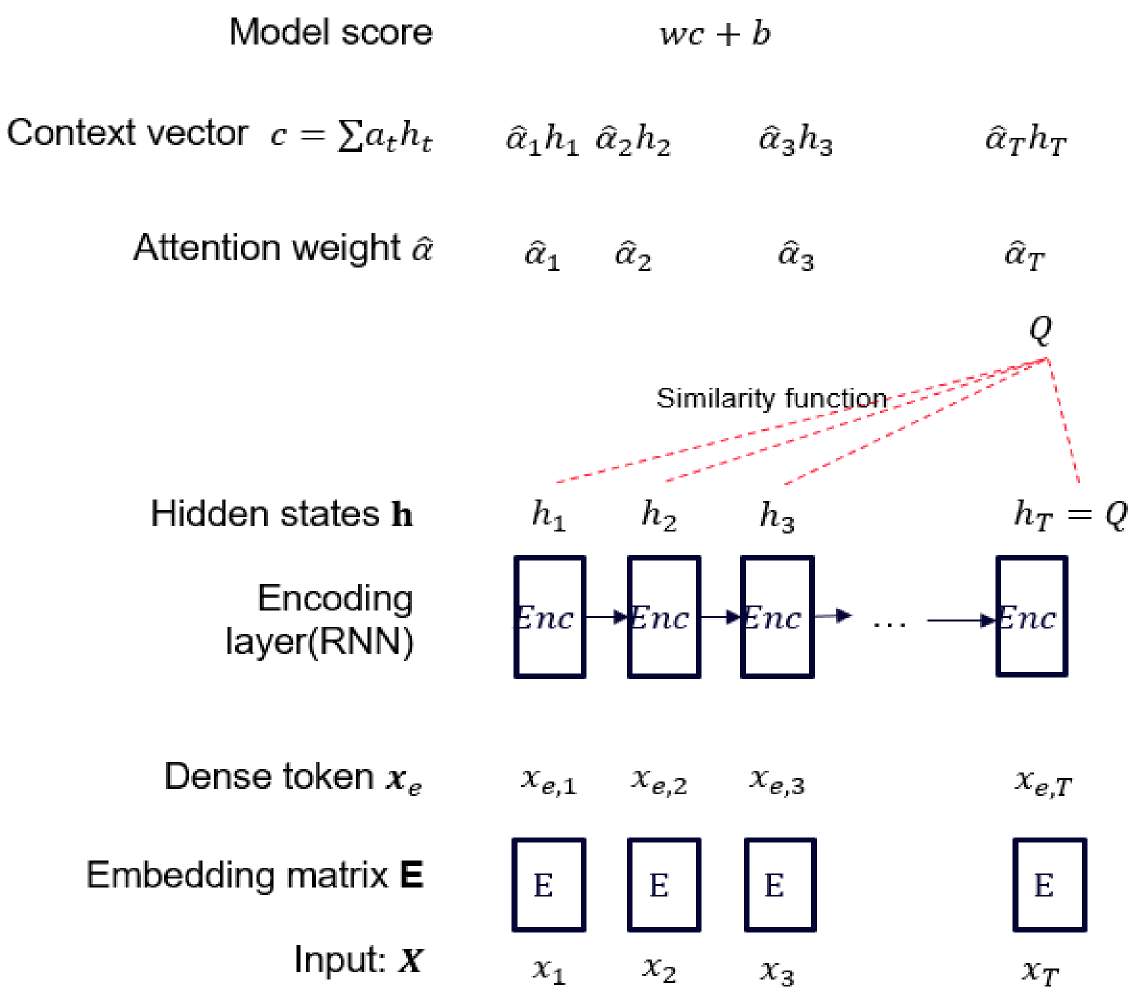
Interpretation

$$\hat{y} = wc + b$$

$$= w\sum a_t h_t + b$$

$$= w(a_1 h_1 + \cdots + a_t h_t) + b$$

$$= w(a_1 x_{e,1} + \cdots + a_t x_{e,t}) + b$$



Preliminary

motivation

3billion

- 동일한 입력 시퀀스를 넣었을 때, 결과는 같은데 attention weight이 다르면 어떻게 될까요?
  (계산을 어찌어찌나오지만, 진짜 해석이 맞는가?)

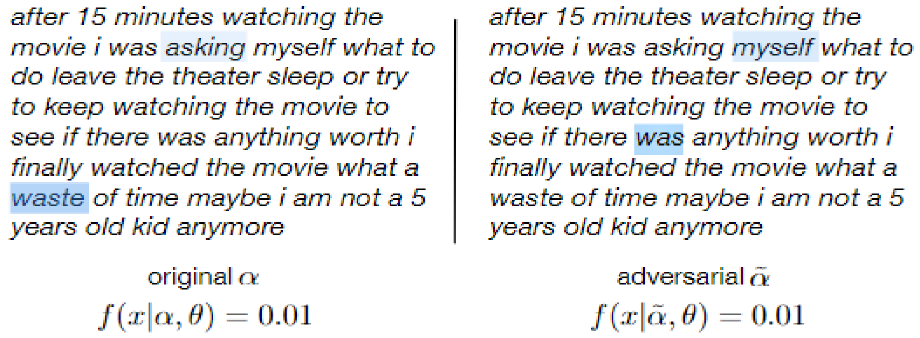


Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model at-

Ref seq:	A	T	A	C	
Input seq:	A	T	**G**	C	
Attention weight1:	0.1	0.1	0.5	0.3	Pathogenicity Prediction: 0.855
Attention weight2:	0.0	0.0	0.3	0.7	Pathogenicity Prediction: 0.855

- 이미 잘 알려진 Explanation 방법론과 Correlation이 없다면 어떨까요?

  예 1) Permutation feature importance와 attention weight의 R^2이 0.5정도밖에 안된다면?
  예 2) Leave one out (LOO)와 attention weight R^2이 0.5정도밖에 안된다면?

1. Introduction

- Attention을 주로 다음과 같이 이용
  - Y(model output) 에 대해 large attention weights을 주는 입력 값을 탐색
  - “Attention provides an important way to explain the workings of neural models“ [1]
  - 많은 연구에서 모델의 해석을 위해서 사용.

- <u>저자:</u>
  - <u>**“Attention을 해석으로 주요 가정은 높은 가중치가 부여된 입력이 모델의 출력에 지대한 영향”**</u>
  - <u>**그런데 과연 이 가정을 평가해본 적 있나?**</u>

- <u>Research questions</u>
  1. Gradient 방법론과 LOO(Leave-one-out)방법론과 상관관계가 있나?
  2. Attention weights가 다르면, 필연적으로 다른 결과값을 나타내는 것인가?

     (매번 attention weights가 다름에도 결과가 동일하다면 = *adversarial attention)*

     => 심지어 permutational attention weigh을 준경우도, 결과가 비슷함. 과연 믿을 수 있나?

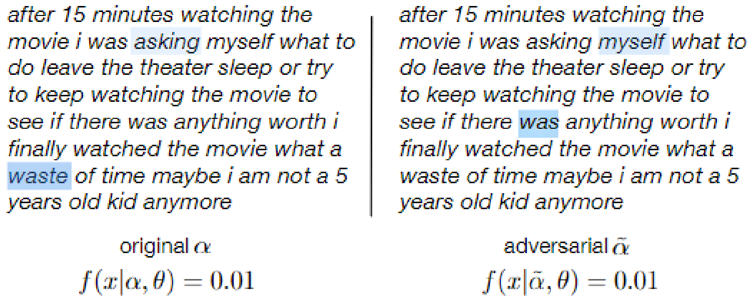


Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

[1] Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. arXiv preprint arXiv:1612.08220

2. Datasets and Task

- **다양한 시퀀스, 테스크에서 Research question의 실험**

1. Binary text classification
    - SST (*Standford Sentiment Treebank*): 1~5점의 ratin인데, (1,2): 부정, (4, 5): 긍정 라벨링을 함.
    - IMDB: sentiment
    - 20 Newsgroups (hockey vs baseball)
    - AG NewCorpus (world vs business)
    - MIMIC ICD9 (*diabetes*): ICU dataset.
        - X= discharge summaries
        - Y= ICD9 code for diabetes
    - MIMIC ICD9 (Chronic vs Acute Anemia)
        - X= discharge summaries (patient with anemia)
        - Y=acute vs chronic

2. Question Answering
- *CNN News articles*: '뉴스단락(본문)'->'질문'->'답변'
- *bAbI:*
    - *Task 1)* single supporting fact
    - *Task 2)*

3. NLI(Natural language infenrece)
- *SNLI* dataset:
    - Entailment, contradictory, neutral 관계를 분류

3. Experiments

Q1. Attention weight ≅ alternative measures (e.g. feature importance)?

- alternative measures: Leave one out, gradient based
    - Kendall-τ correlation

Q2. Attention weight이 다르다면, 예측치도 많이 변하는가?

- counterfactual attention을 생성. (=adversarial attention)

- Change of output distribution: Total Variation Distance (TVD)
- Quantify the difference b/t tow attention distribution: Jensen-Shannon divergence

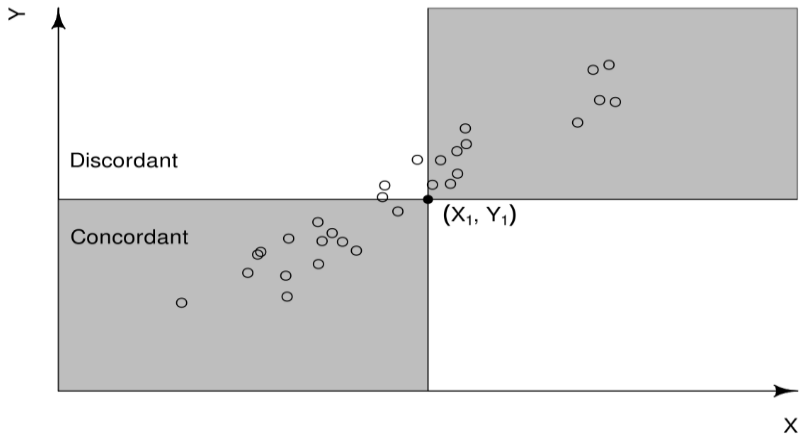
3. Experiments

- Kendall-τ correlation

Eq 1. Kendall rank correlation

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

Figure 1. Kendall rank correlation



- Change of output distribution: Total Variation Distance (TVD)
  - $TVD(\hat{y}_1, \hat{y}_2) = \frac{1}{2}\sum_{i=1}^{|y|} |\hat{y}_{1i} - \hat{y}_{2i}|$
- Quantify the difference b/t tow attention distribution
  - Jensen-Shannon divergence
  - $\text{JSD}(\alpha_1, \alpha_2) = \frac{1}{2}KL[\alpha_1 || (\alpha_1 + \alpha_2)/2] + \frac{1}{2}KL[\alpha_2 || (\alpha_1 + \alpha_2)/2]$

- Correlation b/t attention Feature Importance measure (**Algorithm 1**)
  - Feature importance: Gradient based τ_g
  - Model output change: LOO τ_{loo}
    - Disconnect the computation graph at attention module (not flow through this layer)

Algorithm 1 Feature Importance Computations

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

3. Experiments: result 1

3billion

Table2

1. gradient 방법론과는 moderate ~ high correlation
2. BiLSTM의 경우는 correlation이 0.5을 못 넘음

- BiLSTEM: Bidrectional LSTM
- Average: Linear -> Relu-> Average 하는 엠베딩 방식

		Gradient (BiLSTM) τ_g		Gradient (Average) τ_g		Leave-One-Out (BiLSTM) τ_{loo}	
Dataset	Class	Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig. Frac.	Mean ± 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; note that this largely depends on input length, as correlation does tend to exist, just weakly. Encoders are denoted parenthetically. These are representative results; exhaustive results for all encoders are available to browse online.

3. Experiments: result 2

- Result 2

  => 오히려 Gradient와 LOO와의 상관성이 더 높다

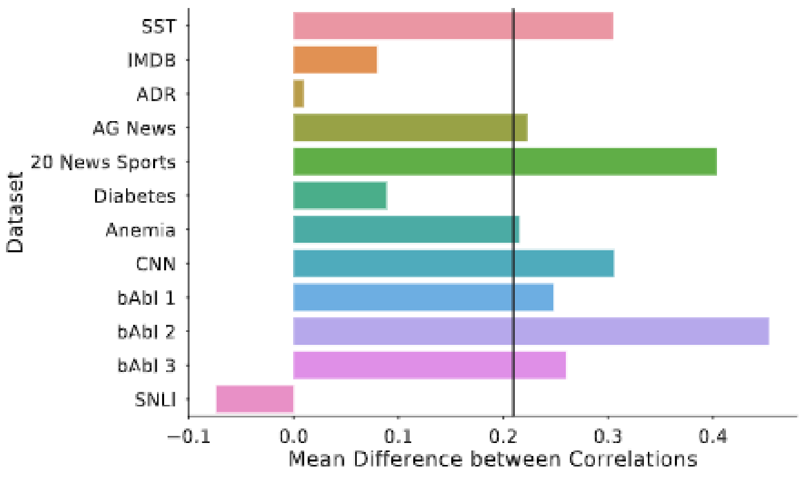


Figure 3: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. LOO scores using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by $>0.2\ \tau_{loo}$.

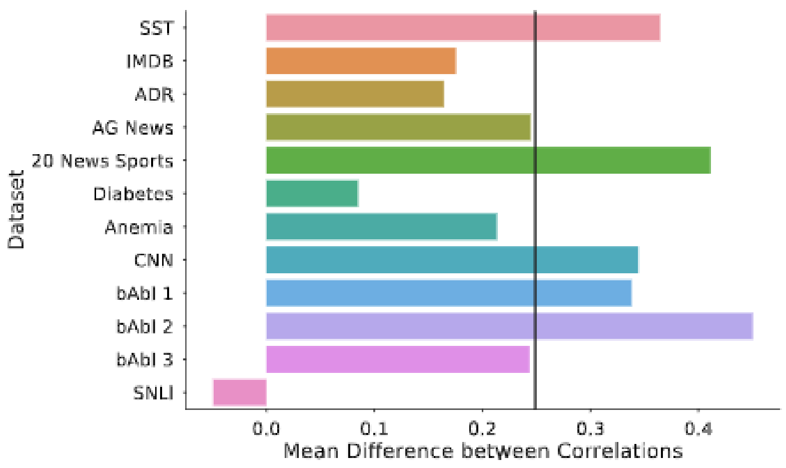


Figure 4: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. Gradients using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by $\sim 0.25\ \tau_g$.

3. Experiments

3billion

- Attention weight을 shuffling하는 경우 어느정도의 예측값이 달라지나?
- Attention weight가 큰 경우, shuffling 하면 예측값이 많이 달라지겠지?

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**

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

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

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

end for

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

3. Experiments : result3

- 일단, attention이 변경되는 경우 결과가 얼마나 바뀌는지 확인해보자.
  가정 "Attention 최대값이 높으면, 결과도 크게바뀔 것"
- X축 model output 차이
  - Max Attention 이 큼에도 불구하고, output difference도 크지 않다
- Y축 원본 attention의 최대값

Predicted as positive
Predicted as negative

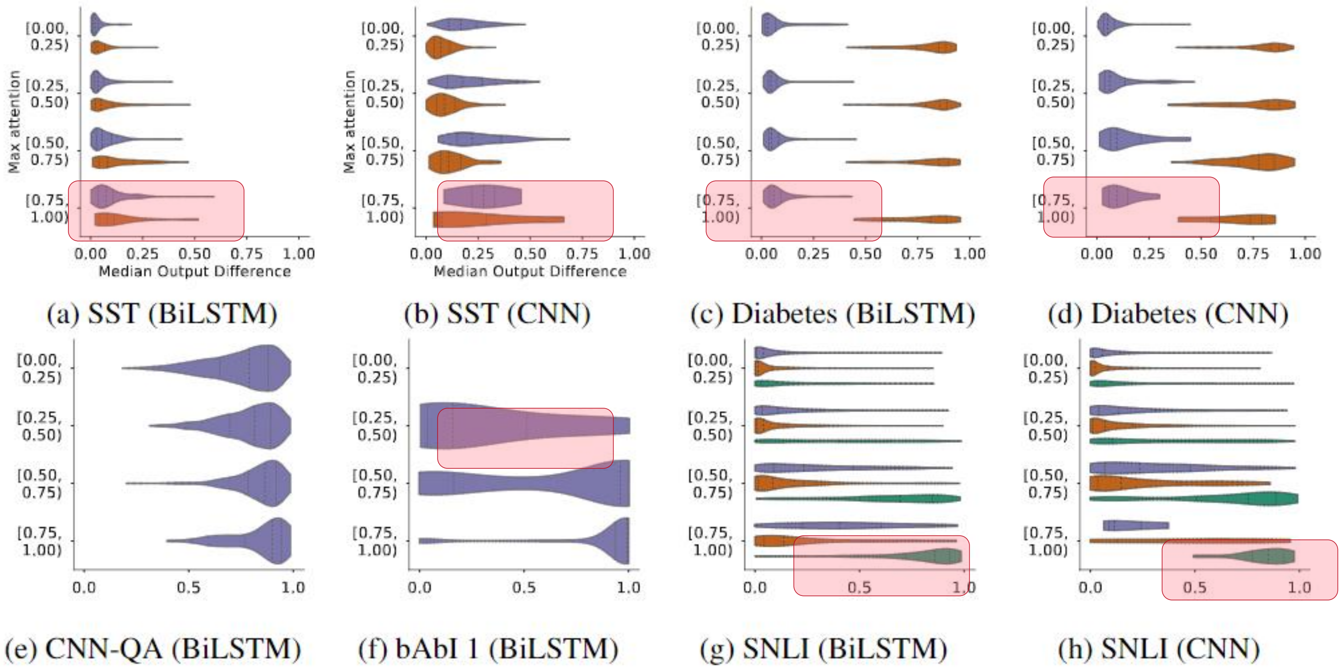


Figure 6: **Median change in output** $\Delta\hat{y}^{med}$ (x-axis) densities in relation to the **max attention** ($\max\hat{\alpha}$) (y-axis) obtained by randomly permuting instance attention weights. Encoders denoted parenthetically. Plots for all corpora and using all encoders are available online.

3. Experiments: result4

- Adversarial attention weight: Prediction결과는 그대로, attention weight은 다른
- Adversarial attention weigh을 찾고, 원래의 분포랑 어느정도 차이나는지 확인해보자.
  - 그래도 분포가 비슷하면 … attention weight으로 해석 인정.

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

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**
    $\hat{y}^{(i)} \leftarrow \mathrm{Dec}(\mathbf{h},\alpha^{(i)})$ ▷ $\mathbf{h}$ 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 $\leftarrow \max_i \mathbb{1}[\Delta\hat{y}^{(i)} \leq \epsilon]\Delta\alpha^{(i)}$

3. Experiments: result4

- **Figure 8:** 결과가 그대로인데, att wt 분포가 다른 경우가 꽤 많더라. 결과에 대한 잘못된 해석을 할 수도 있다.

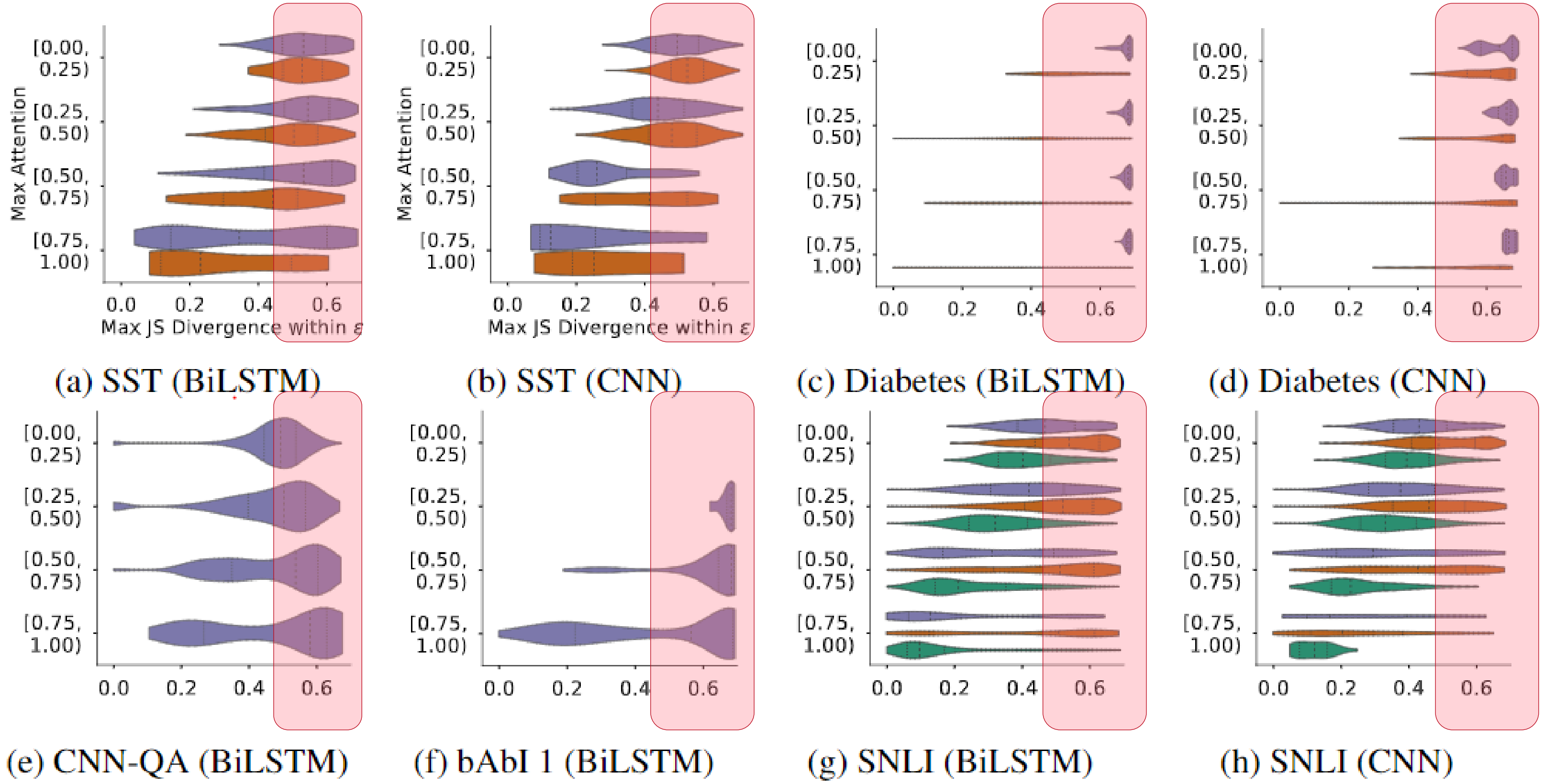


Figure 8: Densities of **maximum JS divergences** (**ϵ-max JSD**) (x-axis) as a function of the **max attention** (y-axis) in each instance for obtained between original and adversarial attention weights.

4. Discussion

- Feature importance와 Attention의 상관성이 약함을 제시함.
- Counterfactual attention weight을 해봐도 동일한 결과가 나오니, 해석시 각각의 해석이 발생할 수 있다.
  - NLP task에서 투명한 모델로 가는데 있어선 최선이지만…

3billion

Your one answer

Thank You

Web. 3billion.io | **Order.** portal.3billion.io | **Email.** support@3billion.io