Attention is ~~not~~*
~~not not~~**
maybe explanation

* Sarthak Jain & Byron C. Wallace, 2019
** Sarah Wiegreffe & Yuval Pinter, 2019

Jacob Danovitch

November 19, 2019

Carleton University

Agenda

RAVENS

Introducing Attention

Attention is not Explanation

Attention is not not Explanation

Other considerations (Time permitting)

- On Identifiability in Transformers
- Is Attention Interpretable?

Discussion

Introducing Attention

What is Attention?

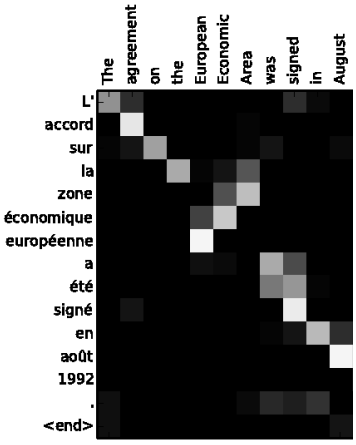


Figure 1: Proposed for neural machine translation as a method to align sequences. [2]

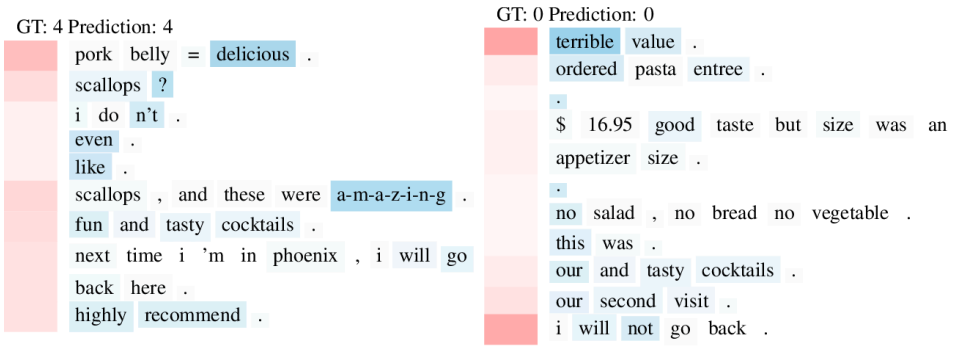


Figure 2: Later proposed for *interpretable* document classification [9].

An input sequence $(x_1, x_2, ..., x_t)$ is passed through a function $\phi(x, W)$ which maps $x_i \rightarrow \alpha_i$. We call these attention weights.

This produces a context vector by taking the sum $\sum_i \alpha_i \cdot x_i$. In essence, we learn to take a weighted average of the input.

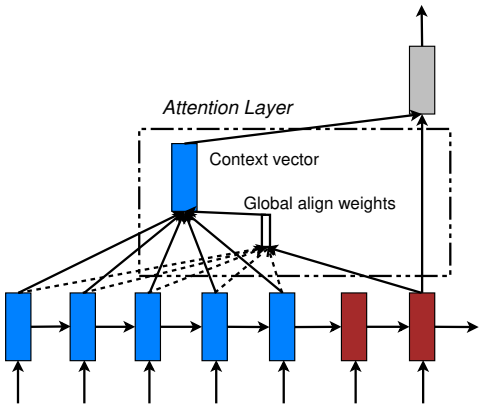


Figure 3: Attention mechanism. [6]

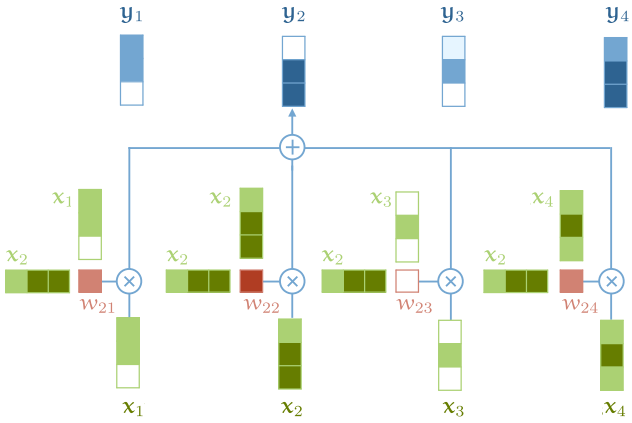


Figure 4: Transformers from Scratch [3].

Successful Applications of Attention

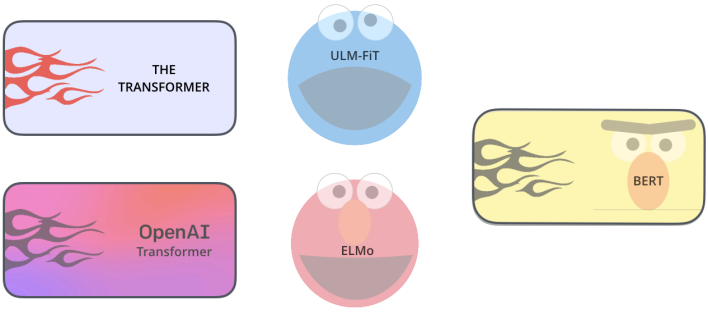


Figure 5: The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning). [1]

Attention is not Explanation

Sarthak Jain and Byron C. Wallace, NAACL-HLT (2019). [5]

Attention is not Explanation

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

Hypotheses:

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

RAVENS

Data: Common NLP benchmarks like IMdB, 20 News Groups, SST, etc.

Notation: A model $\boldsymbol{f} : \mathbb{Z}^{T \times |V|} \rightarrow \mathbb{R}^{T \times D}$ (embedding $\rightarrow$ encoder $\rightarrow \boldsymbol{h}^t$); a similarity function $\phi(\boldsymbol{h}^t, \boldsymbol{Q})$ which outputs an attention distribution $\hat{\boldsymbol{\alpha}}$; and a classification layer θ.

Experiment #1: Feature Importance Correlation

RAVENS

Research question: Do attention weights correlate ($\mathcal{T}$) with feature importance measures?

Measures:

- **Gradient-based methods** (Total Variation Distance, Jensen-Shannon Divergence) $= \mathcal{T}_g$.
  - *"The gradients of the model's output probabilities [...] literally describe the model's decision boundary."* [7]
  - How much do gradients change as a function of input, keeping $\hat{\alpha}$ fixed?
- **Leave-one-out** $= \mathcal{T}_{\text{loo}}$.
  - *"A word's importance can be measured by the difference in model confidence before and after that word is removed from the input."* [4]

Experiment #1: Feature Importance Correlation

RAVENS

Short answer: Not really ($\tau_g = 0.29$). ✗

Experiment #2: Counterfactual Attention Weights

RAVENS

Attention Permutation: Randomly shuffling the elements of $\hat{\boldsymbol{\alpha}}$.

Adversarial Attention: Maximally perturbing $\hat{\boldsymbol{\alpha}}$ while maintaining the same prediction (within some ϵ of $\hat{y}$).

Experiment #2: Counterfactual Attention Weights

RAVENS

Attention Permutation: Authors were able to randomly permute attention weights without significantly affecting output. ✗

Adversarial Attention: Authors were able to perturb $\hat{\boldsymbol{\alpha}}$ (within some ϵ) without significantly affecting output. ✗

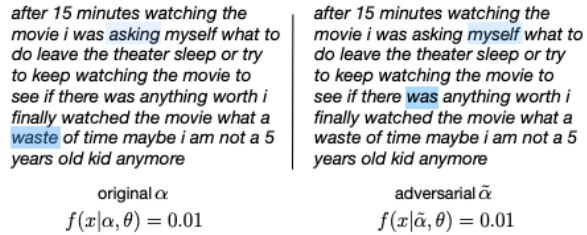


Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

Conclusions

RAVENS

		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; 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.

Conclusions

RAVENS

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.**

One month later...

Attention is not not Explanation

Sarah Wiegreffe and Yuval Pinter (2019). [8]

Attention is not not Explanation

RAVENS

Raises the issues:

- ”Explanation” is ambiguous
- Correlation studies are insufficient
- Adversarial attention experiments had *”little to no meaning”*

Issue # 1: Correlation metric

Kendall-tau correlation measures *rank-based* correlation, which penalizes long-tailed soft attention distribution (where small differences may be negligible in weight, but may affect ranking).

Attention is not not Explanation

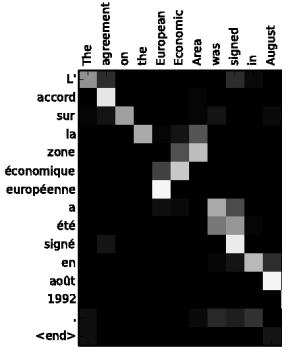
RAVENS

Issue # 2: Adversarial attention

Attention does not provide *the* explanation; it provides *an* explanation. A prediction unaffected by a counterfactual/perturbed attention distribution may simply have multiple valid explanations. ⋆

Issue # 3: Defining explanation

1. What Wallace cares about → "*Attention as a sanity check*: an idea [of] which words in the source text 'should' map to which words in the target text, and it would be a neat demo if a component in the model shows us exactly the patterns we expect."



Issue # 3: Defining explanation

1. What Wallace cares about → "*Attention as a sanity check*: an idea [of] which words in the source text 'should' map to which words in the target text, and it would be a neat demo if a component in the model shows us exactly the patterns we expect."
2. What Pinter cares about → "*Attention as a tool*: the model [...] tells us through attention which part of the text caused it to make the prediction."

Response from Wallace

RAVENS

Jain & Wallace: Interested in *local* explanation. Are attention heatmaps useful for interpretation of individual instances?

Pinter & Wiegreffe: Interested in *global* explanation. Do there exist globally adversarial models?

Response from Wallace

RAVENS

Point of emphasis: *“Attention is not explanation” in the same way that “correlation is not causation.”*

Other considerations

⋆ Recall the possibility of multiple ”attention-explanations” per instance. If many different ”explanations” lead to the same conclusion, would that diminish the value of each explanation?

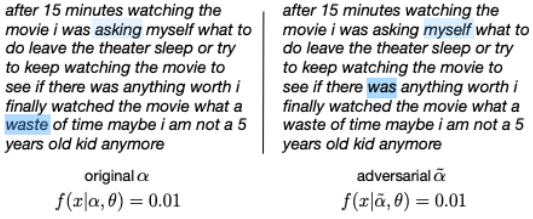


Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

It can be shown that there are ∞ sets of attention weights that lead to the **same prediction**!

$$y = \overbrace{\phi(x, W)}^{A} \cdot x$$

Simplified: We say that some attention weights A are *identifiable* if they map a given $x_i \rightarrow y_i$ *uniquely*. For any $\widetilde{A}_j \in \mathrm{null}(x)$, $(A + \widetilde{A}_j)x = Ax$. When $|x| > \dim(x)^*$, there are ∞ unique $\widetilde{A}_j$ [rank-nullity theorem], and attention is not identifiable.

* Not a problem for normal attention, but is for variants.

Also interesting is **token identifiability**: *"the existence of a one-to-one mapping between contextual embeddings and their corresponding input tokens."*

In other words: Do we take it for granted that token [*embeddings*] maintain their "identity" following attention?

This cannot be proved analytically, so we do within-passage **KNN** using simple MLP. Each token ***t*** should remain most similar to itself after attention.

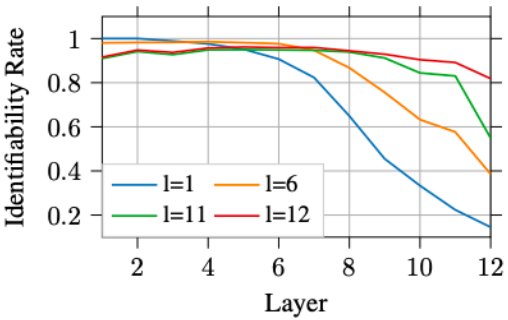


Figure 6: MLP trained on layer l and tested on all layers.

Is Attention Interpretable?

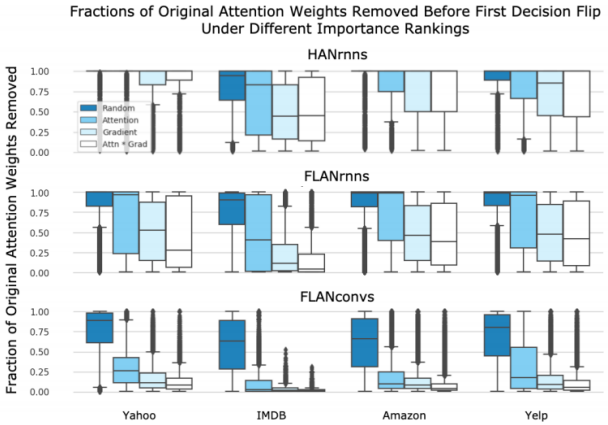


Figure 5: The distribution of fractions of items removed before first decision flips on three model architectures under different ranking schemes. Boxplot whiskers represent the highest/lowest data point within 1.5 IQR of the higher/lower quartile, and dataset names at the bottom apply to their whole column. In several of the plots, the median or lower quartile aren't visible; in these cases, the median/lower quartile is either 1 or very close to 1.

Discussion

In summary:

- [Jain and] **Wallace**: *"Attention is not explanation" in the same way that "correlation is not causation"*
- [Wiegreffe and] **Pinter**: *"Attention might still be explanation"*
- **Serrano and Smith**: attention *"should not be treated as justification for a decision"*
- **Bruner et al.**: *"attention visualizations are misleading"*

Discussion

RAVENS

- Agree or disagree: *“Attention is not explanation” in the same way that “correlation is not causation”*

Discussion

RAVENS

- ~~Agree or disagree:~~
  ~~*“Attention is not explanation” in the same way that “correlation is not causation”*~~
- What does “explanation” mean to you?
  - Can attention be explanation considering the token identification and decision flipping results?

Discussion

RAVENS

- ~~Agree or disagree:~~
  ~~*"Attention is not explanation" in the same way that "correlation is not causation"*~~
- ~~What does "explanation" mean to you?~~
  - ~~Can attention be explanation considering the token identification and decision flipping results?~~
- What could attention be measuring, if not importance to the output?

[1] J. Alammar.
The illustrated bert, elmo, and co. (how nlp cracked transfer learning), 2019.

[2] D. Bahdanau, K. Cho, and Y. Bengio.
Neural machine translation by jointly learning to align and translate.
CoRR, abs/1409.0473, 2014.

[3] P. Bloem.
Transformers from scratch, 2019.

[4] S. Feng, E. Wallace, A. Grissom II, M. Iyyer, P. Rodriguez, and J. Boyd-Graber.
Pathologies of neural models make interpretations difficult.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics.

[5] S. Jain and B. C. Wallace.
Attention is not explanation, 2019.

[6] T. Luong, H. Pham, and C. D. Manning.
Effective approaches to attention-based neural machine translation.
Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015.

[7] A. S. Ross, M. C. Hughes, and F. Doshi-Velez.
Right for the right reasons: Training differentiable models by constraining their explanations.
Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Aug 2017.

[8] S. Wiegreffe and Y. Pinter.
Attention is not not explanation.
Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[9] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy.
Hierarchical attention networks for document classification.
In *HLT-NAACL*, 2016.