Is Attention Interpretable?

Sofia Serrano∗ Noah A. Smith∗†

∗Paul G. Allen School of Computer Science & Engineering,

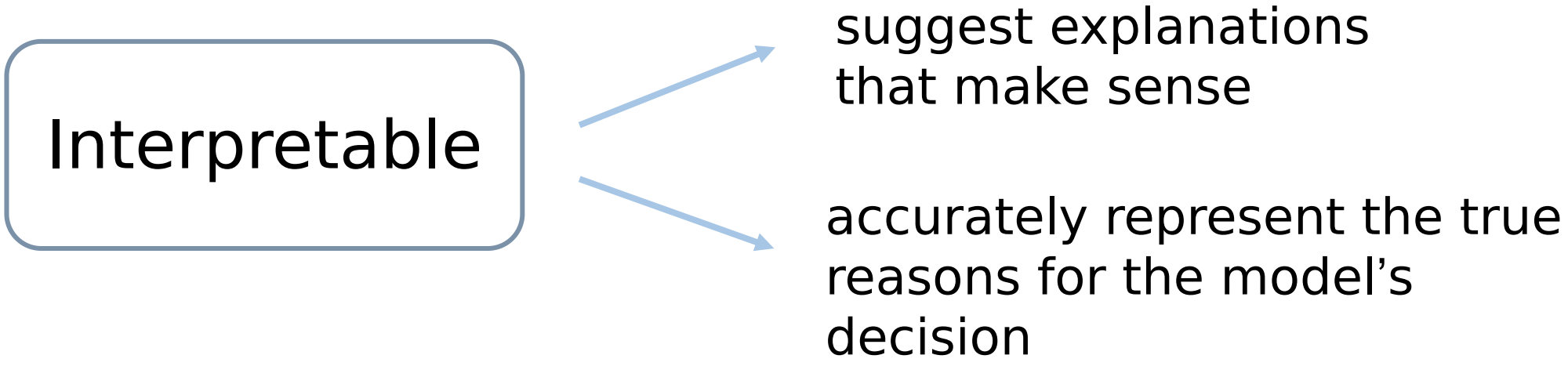
University of Washington, Seattle, WA, USA

†Allen Institute for Artificial Intelligence, Seattle, WA, USA

ACL 2019

Motivation

- explore whether attention mechanisms can identify the relative importance of inputs to the full model



Intermediate Representation Erasure

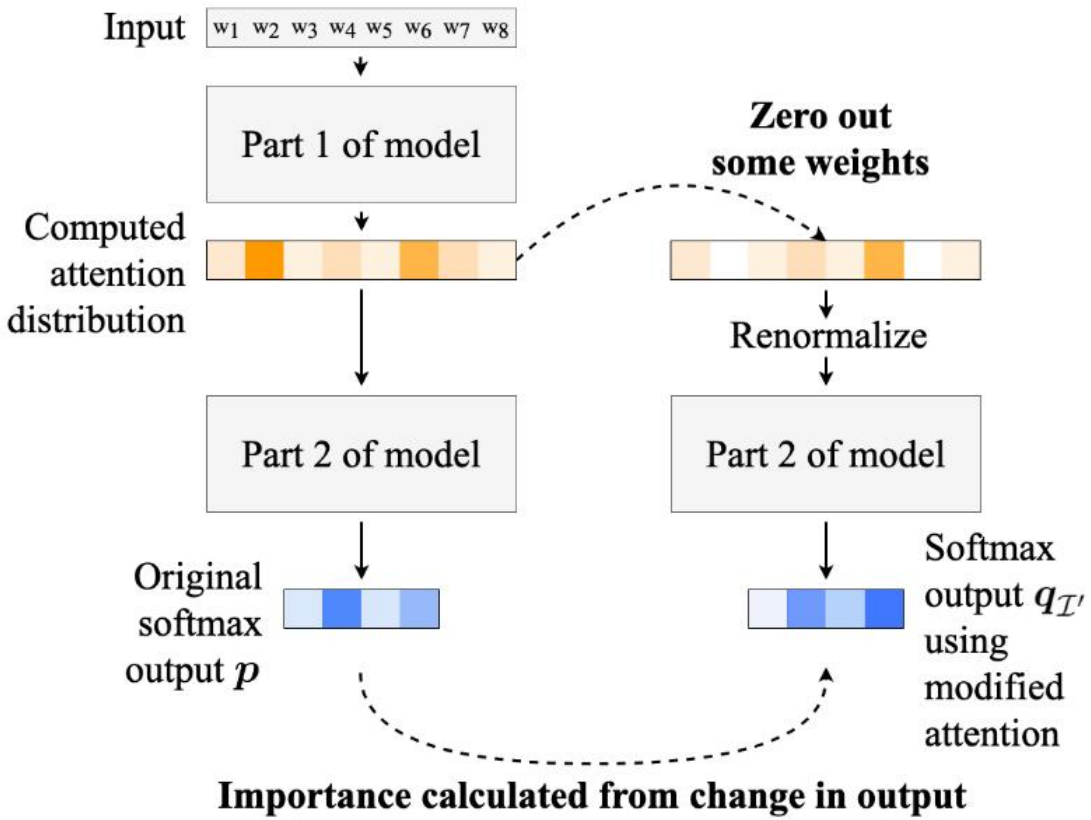


Figure 1: Our method for calculating the importance of representations corresponding to zeroed-out attention weights, in a hypothetical setting with four output classes .

Data

Dataset	Av. # Words	(s.d.)	Av. # Sents.	(s.d.)	# Train. + Dev.	# Test	# Classes
Yahoo Answers	104	(114)	6.2	(5.9)	1,400,000	50,000	10
IMDB	395	(259)	16.2	(10.7)	122,121	13,548	10
Amazon	73	(48)	4.3	(2.6)	3,000,000	650,000	5
Yelp	125	(109)	7.0	(5.6)	650,000	50,000	5

Table 1: Dataset statistics.

Model

- Two layers of attention:
  - first to the word tokens in each sentence;
  - then to the resulting sentence representations.
- Attention calculation

$$\mathbf{u}_i = \tanh(\mathbf{W}_\ell \mathbf{h}_i + \mathbf{b}_\ell)$$

$$\alpha_i = \frac{\exp \mathbf{u}_i^\top \mathbf{c}_\ell}{\sum_i \exp \mathbf{u}_i^\top \mathbf{c}_\ell}$$

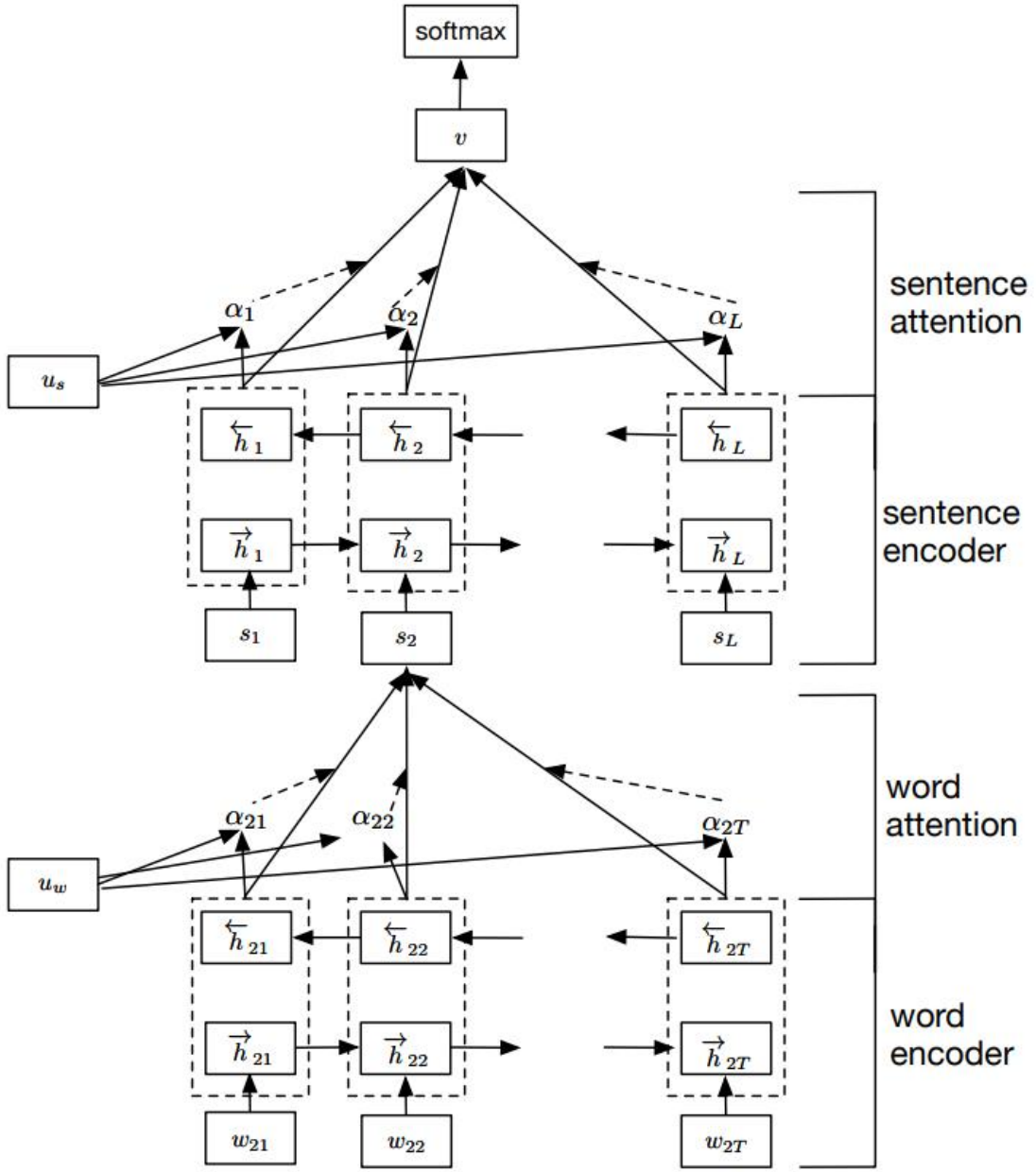


Figure 2: Hierarchical Attention Network.

Single Attention Weights' Importance

$$\Delta \mathrm{JS} = \mathrm{JS}(\boldsymbol{p}, \boldsymbol{q}_{\{i^*\}}) - \mathrm{JS}(\boldsymbol{p}, \boldsymbol{q}_{\{r\}})$$

$$\Delta\alpha = \alpha_{i^*} - \alpha_r$$

i^* : the component with the highest attention

α_{i^*} : attention of i^* .

r : random attended item

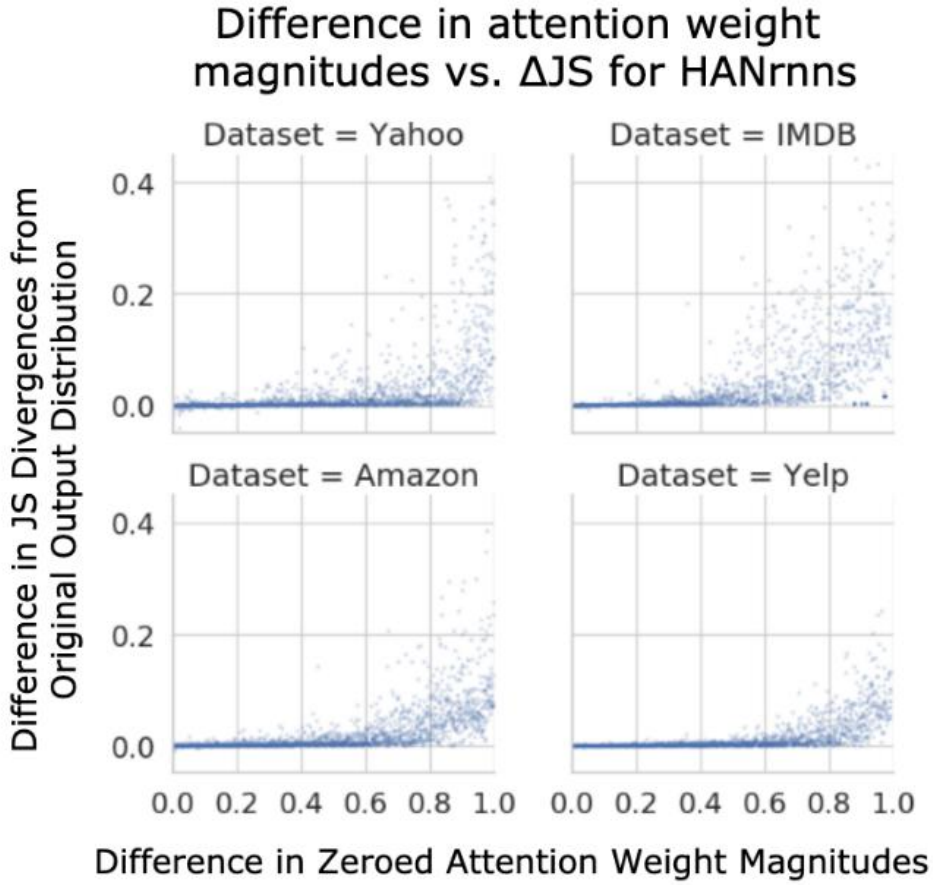


Figure 3: Difference in attention weight magnitudes versus ΔJS for HANrnns, comparable to results for the other architectures; for their plots, see Appendix A.2.

Single Attention Weights' Importance

- in the vast majority of cases, erasing i^* does not change the decision ("no" row of each table)
- the difference in impacts between i^* and r is almost identical (i.e., ΔJS values close to 0 or the many cases where neither i^* nor r cause a decision flip)

Remove random: Decision flip?

Remove i^*: Decision flip?

Yahoo

	Yes	No
Yes	0.5	8.7
No	1.3	89.6

IMDB

	Yes	No
Yes	2.2	12.2
No	1.4	84.2

Amazon

	Yes	No
Yes	2.7	7.6
No	2.7	87.1

Yelp

	Yes	No
Yes	1.5	8.9
No	1.9	87.7

Table 2: Percent of test instances in each decision-flip indicator variable category for each HANrnn.

Importance of Sets of Attention Weights

- erase representations from the top of the ranking downward until the model's decision changes.
- Alternative Importance Rankings
  - randomly rank importance
  - gradient
  - gradient * attention

Importance of Sets of Attention Weights

- discover much smaller decision-flipping sets of items than attention weights.
- we should be skeptical of trusting groups of attention weight magnitudes as importance indicators.
- Attention Does Not Optimally Describe Model Decisions

Takeaways

- Looking at attention distribution can be misleading
- Can wrongly imply that
  - A small number of representations are responsible for the decision
  - Some items are more important than others that are actually more influential to the model decision
- Depending on the model structure preceeding the attention layer, attention weights might be much worse at describing importance

Open questions

- How to move past decision changes as signal of importance?
  - Would allow analysis of tasks with structured outputs
  - Would enable testing other less strict definitions of importance
- Does attention function differently depending on its formulation or its location in a model?
- What othe things might attention possibly tell us?