Supervising Model Attention with Human Explanations for Robust Natural Language Inference

by Joe Stacey[1], Yonatan Belinkov[2] and Marek Rei[1]
[1]Imperial College London, [2]Technion - Israel Institute of Technology
Contact: j.stacey20@imperial.ac.uk
GitHub: https://github.com/joestacey/NLI_with_a_human_touch

1. **We train with human explanations to create more robust NLI models**, paying more attention to the words that humans think are important.

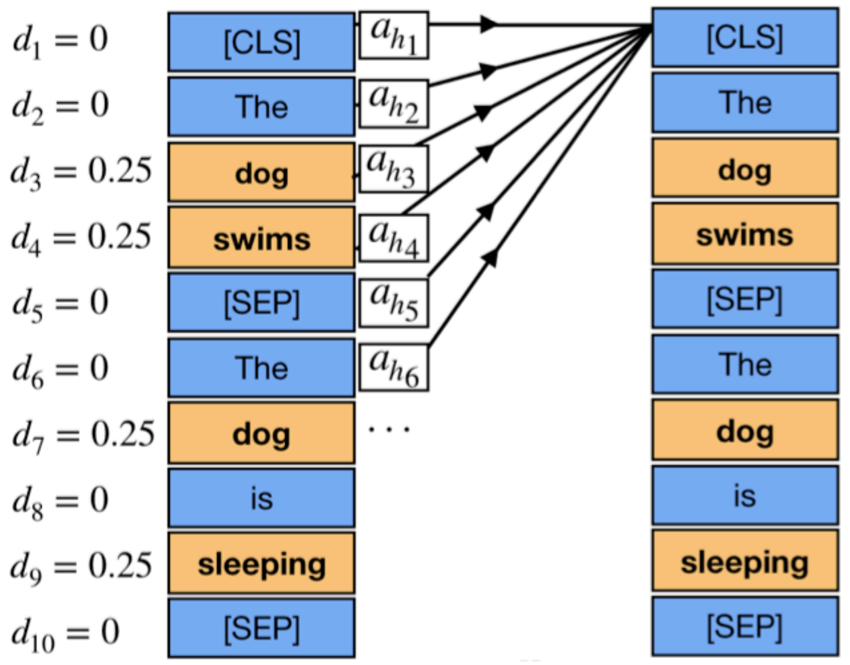
e-SNLI human explanations:

Premise:
*Wet brown **dog swims** towards camera.*

Hypothesis:
*A **dog** is **sleeping** in his bed.*

Explanation for contradiction class:
*A **dog** cannot be **sleeping** while he **swims**.*

2. **We achieve this by creating a desired attention distribution,** based on the content of the e-SNLI human explanations.



3. **An auxiliary loss** encourages more attention to this distribution, supervising either: 1) existing attention heads, or 2) an additional attention layer.

$$a_{h_i} = \frac{\exp\left(q_{h_{CLS}}^T k_{h_i} / \sqrt{d_k}\right)}{\sum_{j=1}^{n} \exp\left(q_{h_{CLS}}^T k_{h_j} / \sqrt{d_k}\right)}$$

$$Loss_{Total} = Loss_{NLI} + \frac{\lambda}{H} \sum_{h=1}^{H} \sum_{i=1}^{n} (a_{h_i} - d_i)^2$$

4. **We see significant improvements** in ID and OOD performance from both approaches.

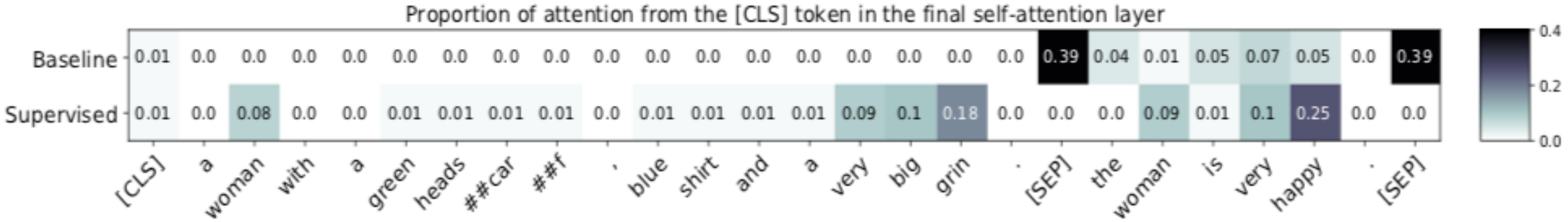
Performance compared to our BERT baseline:

	Dev	Test	Hard	MNLI mi	MNLI ma	ANLI	HANS
BERT baseline	90.05	89.77	79.36	72.52	72.28	31.81	56.83
Ours (extra layer)	90.40	90.09	79.96	73.03	73.10	31.47	57.85
Improvement	**+0.35**†‡	**+0.32**†‡	**+0.60**†‡	**+0.51**†	**+0.82**†‡	-0.34	+1.02
Ours (existing attention)	90.45	90.17	80.15	73.36	73.19	31.41	58.42
Improvement	**+0.40**†‡	**+0.40**†‡	**+0.79**†‡	**+0.84**†‡	**+0.91**†‡	-0.40	**+1.59** †

Improvements compared to prior work:

	SNLI	Δ	MNLI	Δ	SNLI-hard	Δ	Params.
BERT Baseline	89.77		72.40		79.36		109m
LIREx-adapted	**90.79**	**+1.02**†	71.55	-0.85†	79.39	+0.03	453m
Pruthi et al-adapted.	89.99	+0.22†	73.27	+0.87†	79.90	+0.54†	109m
Ours (extra layer)	90.09	+0.35†	73.06	+0.67†	79.96	+0.60†	109m
Ours (existing attention)	90.17	+0.40†	**73.28**	**+0.88**†	**80.15**	**+0.79**†	109m

5. **More attention is paid to the premise,** even in the layers before the supervised layer, helping to mitigate the hypothesis-only bias. We also see less attention paid to stop-words and more attention paid to nouns, verbs and adjectives.



Proportion of attention to the premise:

Baseline attention to premise*: **22.86%**

Supervised attention to premise*: **50.89%**

* Attention to premise and 1st [SEP] token

Proportion of attention by PoS tag:

PoS Tag	12 heads	3 heads	Baseline
Noun	**54.3**	43.5	28.1
Verb	**20.4**	18.2	14.3
Adjective	**8.9**	8.3	5.2
Adposition	4.1	5.0	**7.8**
Determiner	3.4	6.0	**14.3**
Punctuation	0.9	7.7	**14.2**
Auxiliary	0.9	3.1	**8.2**
Other	7.1	8.2	7.9

Most attended to words:

Baseline		**Supervised**	
Words	%	Words	%
.	18.0	man	2.7
a	5.2	outside	2.5
is	4.0	woman	1.7
are	2.6	people	1.7
the	2.5	sitting	1.5