Do Transformer Models Show Similar Attention Patterns to Task-Specific Human Gaze?

Oliver Eberle[*,1], **Stephanie Brandl**[*,1,2], **Jonas Pilot**[1], **Anders Søgaard**[2]

[*]equal contribution
[1]Machine Learning Group, TU Berlin, Germany
[2]University of Copenhagen, Denmark

eMail: oliver.eberle@tu-berlin.de, brandl@di.ku.dk

Introduction

Learned self-attention functions in state-of-the-art NLP models often correlate with human attention. We investigate whether self-attention in large-scale pre-trained language models is as predictive of human eye fixation patterns during task-reading as classical cognitive models of human attention. We compare attention functions across two task-specific reading datasets for sentiment analysis and relation extraction. We find the predictiveness of large-scale pre-trained self-attention for human attention depends on 'what is in the tail', e.g., the syntactic nature of rare contexts. Further, we observe that task-specific fine-tuning does not increase the correlation with human task-specific reading. Through an input reduction experiment we give complementary insights on the sparsity and fidelity trade-off, showing that lower-entropy attention vectors are more faithful.

Experiments

We compare attention functions for a variety of computational models with the task-specific eye-tracking recordings from ZuCo [1]: 12 participants reading sentences from the English Wikipedia (relation extraction) and SST (sentiment reading):

Type of model	Model(s)	Attention
Human	Eye-tracking	total fixation times across participants
Transformers	(fine-tuned) BERT (base & large), RoBERTa, T5	attention flow (from different layers) [2], mean across last (raw) attention layer
Shallow	CNN & single-self-attention	layer-wise relevance propagation
Cognitive	E-Z Reader [3]	predicted gaze
Frequency	British National Corpus	inverse word frequency

Results

Main results. Spearman correlation on sentence and token-level between aforementioned models and human gaze.

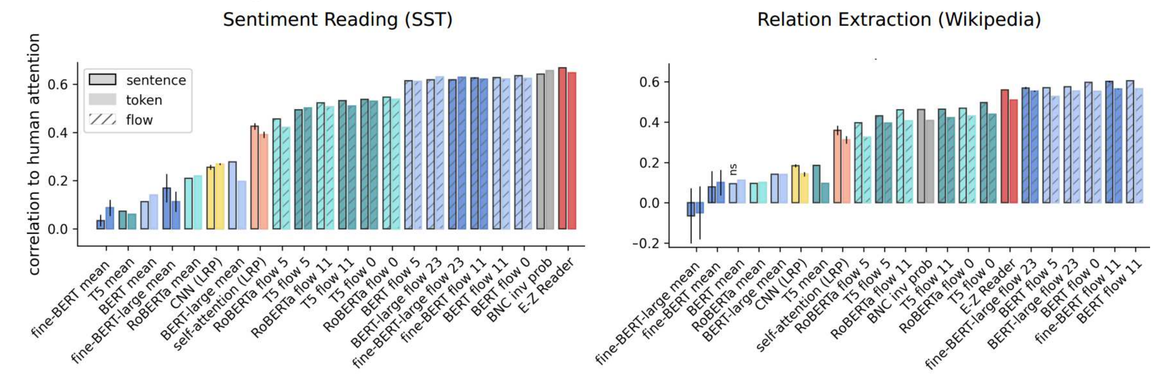

Figure 1: Spearman correlation analysis between human attention and different models for two task settings. Solid bar edges indicate sentence-level correlations in contrast to a token-level analysis. Left: Sentiment Reading on the SST dataset. Right: Relation Extraction on Wikipedia. Standard deviations over five seeds are shown for fine-tuned models and correlations are statistically significant with p ¡ 0.01 unless stated otherwise (ns: not significant).

- E-Z Reader and the frequency baseline on BNC correlate better with human gaze on SST but not in Wikipedia
- fine-tuning and model size does not influence correlation for BERT
- correlation with attention flow does not change across layers
- shallow models correlate much less than Transformers
- mean across last (raw) attention layer does not show high correlations

Correlations based on word predictability. We compare correlations to human fixations with attention flow values for Transformer models in the last layer, the E-Z Reader and the BNC baseline for different word predictability scores (based on 5-gram Kneser-Ney).

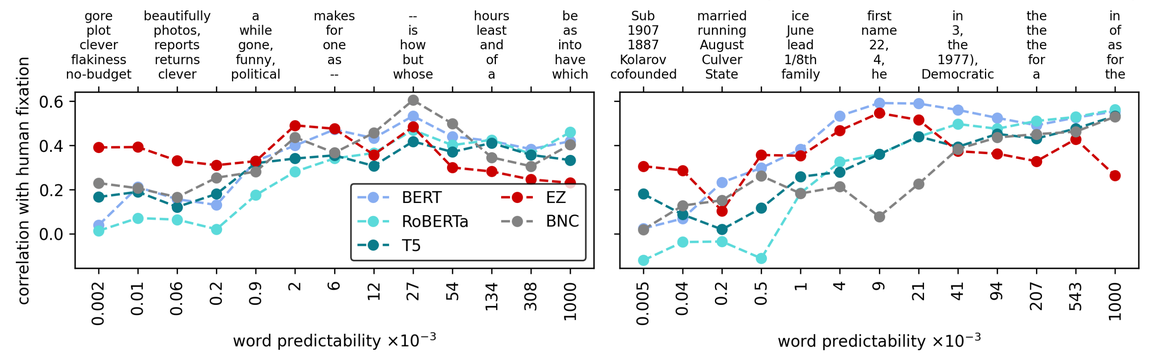

Figure 2: Correlation between human fixations and different models for SST (left) and Wikipedia (right) with respect to word predictability in equally sized bins. Word predictability scores, were calculated with a 5-gram Kneser-Ney language model. Respective bin limits are given on the x-axis. Samples for every other bin are displayed on the upper x-axis.

- Transformer models correlate better for more predictable words on both datasets
- E-Z Reader is less influenced by word predictability
- on SST, Transformers only pass the E-Z Reader on the most predictable tokens (word predictability > 0.03)

Results cont'd

Correlations based on POS tags. We also compare correlations to human fixations based on the top-6 (most tokens) Part-of-speech tags.

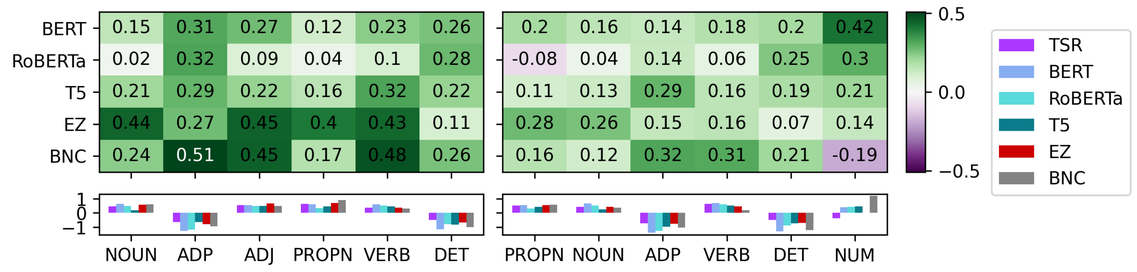

Figure 3: Upper: Correlations between human fixation and different models for SST (left) and Relation Extraction (right) for the six most common POS tags. Lower: Average attention value after standardization (mean=0, std=1) for respective POS tag and model

- on SST, correlations with E-Z Reader are very consistent across POS tags
- attention flow shows weak correlations on proper nouns (0.12), nouns (0.16) & verbs (0.16)
- the BNC frequency baseline correlates well with human fixations on adpositions (ADP) which both assign comparably low values
- proper nouns (PROPN) are overestimated in BNC as a result of their infrequent occurrence

Natural Reading. ZuCo contains a subset of 48 sentences that were presented both in a task-specific and a natural reading setting. This allows for a direct comparison of correlation strength.

Table 1: Correlations between human fixations and models on 48 duplicates from the ZuCo dataset for both natural reading (NR) and relation extraction (task-specific reading - TSR).

	BERT mean	RoBERTa mean	T5 mean	fine-BERT mean	T5 flow 11	RoBERTa flow 11	BNC inv prob	E-Z Reader	fine-BERT flow 11	BERT flow 11	ZuCo NR
NR	.12	.09	.16	.15	.48	.52	.58	.57	.67	.69	-
TSR	.12	.14	.20	.23	.45	.48	.49	.53	.61	.62	.72

Faithfulness and Entropy analysis. We perform a perturbation analysis by unmasking tokens in order from highest to lowest importance in a task-tuned BERT model.

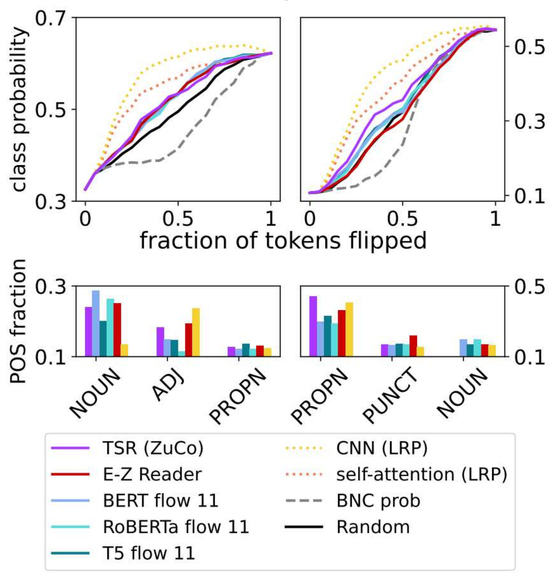

Figure 4: Results of our reduction analysis where most important tokens are selected and fed into fine-tuned BERT models for sentiment classification (left) and relation extraction (right). Upper: we gradually measure output probability for the true label. Higher AUC reflects a stronger model sensitivity to adding important tokens. Lower: Fractions of most-selected POS tags at the first flip are displayed for human attention (TSR), flow 11, E-Z and BNC token probability

	TSR (ZuCo)	E-Z Reader	BNC inv prob	CNN (LRP)	self-attention (LRP)	BERT flow 11	RoBERTa flow 11	T5 flow 11	BERT mean	RoBERTa mean	T5 mean
SR	3.44	3.44	3.40	2.93	2.16	3.57	3.61	3.61	2.37	2.65	2.45
TSR	3.38	3.46	3.39	2.98	1.81	3.54	3.60	3.63	2.48	2.56	2.29

Table 2: Mean entropy over all sentences for each task setting. Lower entropy means sparser token importance. The maximal entropy of a uniform model is 4.09 bits

- shallow, from-scratch trained models (CNN, shallow self-attention) are most faithful and more sparse
- Transformer flow, Humans attention and E-Z Reader are moderately faithful and less sparse.
- human task-specific reading is sub-optimal relative to task-solving, heavily regularized by natural reading patterns
- Sparsity - Faithfulness - Correlation trade-off

Conclusion

In our experiments, we first and foremost found that Transformers, and especially BERT models, are competitive to the E-Z Reader in terms of explaining human attention in task-specific reading. For this to be the case, computing attention flow scores (rather than raw attention weights) is important. Even so, the E-Z Reader remains better at hard-to-predict words and is less sensitive to part of speech. While Transformers thus have some limitations compared to the E-Z Reader, our results indicate that cognitive models have placed too little weight on high-level word co-occurrence statistics. Generally, Transformers and the E-Z Reader correlate much better with human attention than other, shallow from-scratch trained sequence labeling architectures. Our input reduction experiments suggest that in a sense, both pre-trained language models *and* humans have suboptimal, i.e., less sparse, task-solving strategies, and are heavily regularized by what is optimal in natural reading contexts.

References

[1] Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. Scientific data, 5(1), 1-13

[2] Abnar, S., and Zuidema, W. (2020). Quantifying Attention Flow in Transformers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4190-4197).

[3] Reichle, E. D., Pollatsek, A., Fisher, D. L., and Rayner, K. (1998). Toward a model of eye movement control in reading. Psychological review, 105(1), 125.