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]
[1]Machine Learning Group, TU Berlin, Germany [2]University of Copenhagen, Denmark
oliver.eberle@tu-berlin.de, {brandl, soegaard}@di.ku.dk
[*]Authors contributed equally.

Abstract

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.

1 Introduction

The usefulness of learned self-attention functions often correlates with how well it aligns with human attention (Das et al., 2016; Klerke et al., 2016; Barrett et al., 2018; Zhang and Zhang, 2019; Klerke and Plank, 2019). In this paper, we evaluate how well attention flow (Abnar and Zuidema, 2020) in large language models, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020), aligns with human eye fixations during task-specific reading, compared to other shallow sequence labeling models (Lecun and Bengio, 1995; Vaswani et al., 2017) and a classic, heuristic model of human reading (Reichle et al., 2003). We compare the learned attention functions and the heuristic model across two task-specific English reading tasks, namely sentiment analysis (SST movie reviews) and relation extraction (Wikipedia), as well as natural reading, using a publicly available dataset with eye-tracking recordings of native speakers of English (Hollenstein et al., 2018).

Contributions We compare human and model attention patterns on both sentiment reading and relation extraction tasks. In our analysis, we compare human attention to pre-trained Transformers (BERT, RoBERTa and T5), from-scratch training of two shallow sequence labeling architectures (Lecun and Bengio, 1995; Vaswani et al., 2017), as well as to a frequency baseline and a heuristic, cognitively inspired model of human reading called the E-Z Reader (Reichle et al., 2003). We find that the heuristic model correlates well with human reading, as has been reported in Sood et al. (2020b). However when we apply *attention flow* (Abnar and Zuidema, 2020), the pre-trained Transformer models also reach comparable levels of correlation strength. Further fine-tuning experiments on BERT did not result in increased correlation to human fixations. To understand what drives the differences between models, we perform an in-depth analysis of the effect of word predictability and POS tags on correlation strength. It reveals that Transformer models do not accurately capture tail phenomena for hard-to-predict words (in contrast to the E-Z Reader) and that Transformer attention flow shows comparably weak correlation on (proper) nouns while the E-Z Reader predicts importance of these more accurately, especially on the sentiment reading task. In addition, we investigate a subset of the ZuCo corpus for which aligned task-specific and natural reading data is available and find that Transformers correlate stronger to natural reading patterns. We test faithfulness of these different attention patterns to produce the correct classification via an input reduction experiment on task-tuned BERT models. Our results highlight the trade-off between model faithfulness and sparsity when comparing importance scores to human attention, i.e., less sparse (higher entropy) attention vectors tend to be less faithful with respect to model predictions. Our code is available at github.com/oeberle/task_gaze_transformers.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4295 - 4309
May 22-27, 2022 ©2022 Association for Computational Linguistics

2 Pre-trained Language Models vs Cognitive Models

Church and Liberman (2021) discuss how NLP has historically benefited from rationalist and empiricist methodologies, something that holds for cognitive modeling in general. The vast majority of application-oriented work in NLP today relies on pre-trained language models or other large-scale data-driven models, but in cognitive modeling, most approaches remain heuristic and rule-based, or hybrid, e.g., relying on probabilistic language models to quantify surprisal (Rayner and Reichle, 2010; Milledge and Blythe, 2019). This is for good reasons: Cognitive modeling values interpretability (even) more, often suffers from data scarcity, and is less concerned with model reusability across different contexts.

This paper presents a head-to-head comparison of the E-Z Reader and pre-trained Transformer-based language models. We are not the first to evaluate pre-trained language models and large-scale data-driven models as if they were cognitive models. Chrupała and Alishahi (2019), for example, use representational similarity analysis to correlate sentence encodings in pre-trained language models with fMRI signals; Abdou et al. (2019) correlate sentence encodings with gaze-derived representations. More generally, it has been argued that cognitive evaluations are in some cases practically superior to standard evaluation methodologies in NLP (Søgaard, 2016; Hollenstein et al., 2019). We return to this in the Discussion and Conclusion §6.

Commonly, pre-trained language models are disregarded as cognitive models, since they are most often implemented as computationally demanding batch learning algorithms, processing data "at once". Günther et al. (2019) points out that this is an artefact of their implementation, and online learning of pre-trained language models is possible, yet impractical. Generally, several researchers have argued for taking pre-trained language models seriously as cognitive models (Rogers and Wolmetz, 2016; Mandera et al., 2017; Günther et al., 2019). In the last section, §6, we discuss some of the implications of comparisons of pre-trained language models and cognitive models – for cognitive modeling, as well as for NLP. In our experiments, we focus on Transformer architectures that are currently the dominating pre-trained language models and a *de facto* baseline for modern NLP research.

3 Experiments

3.1 Data

The ZuCo dataset (Hollenstein et al., 2018) contains eye-tracking data for 12 participants (all English native speakers) performing natural reading and relation extraction on 300 and 407 English sentences from the Wikipedia relation extraction corpus (Culotta et al., 2006) respectively and sentiment reading on 400 samples of the Stanford Sentiment Treebank (SST) (Socher et al., 2013). For our analysis, we extract and average word-based total fixation times across participants and focus on the task-specific relation extraction and sentiment reading samples.

3.2 Models

Below we briefly describe our used models and refer to Appendix A for more details.

Transformers The superior performance of Transformer architectures across broad sets of NLP tasks raises the question of how task-related attention patterns really are. In our experiments, we focus on comparing task-modulated human fixations to attention patterns extracted from the following commonly used models: (a) We use both pre-trained uncased BERT-base and large models (Devlin et al., 2019) as well as fine-tuned BERT models on the respective tasks. BERT was originally pre-trained on the English Wikipedia and the BookCorpus. (b) The RoBERTa model has the same architecture as BERT and demonstrates better performance on downstream tasks using an improved pre-training scheme and the use of additional news article data (Liu et al., 2019). (c) The Text-to-Text Transfer Transformer (T5) uses an encoder-decoder structure to enable parallel task-training and has demonstrated state-of-the-art performance over several transfer tasks including sentiment analysis and natural language inference (Raffel et al., 2020).

We evaluate different ways of extracting token-level importance scores: We collect attention representations and compute the mean attention vector over the final layer heads to capture the mixing of information in Transformer self-attention modules as in Hollenstein and Beinborn (2021) and present this as *mean* for all aforementioned Transformers.

To capture the layer-wise structure of deep Transformer models we compute attention flow (Abnar and Zuidema, 2020). This approach considers the

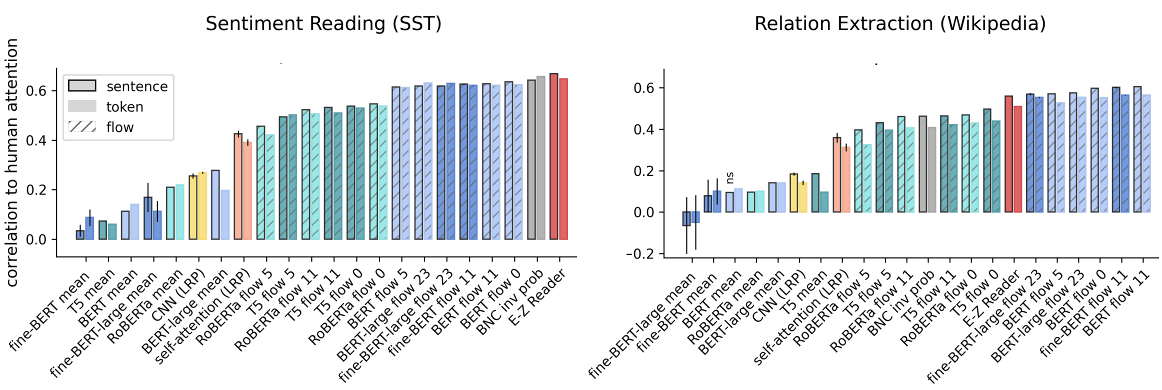

Figure 1: Spearman correlation analysis between human attention and different models for two task settings. Solid bar edges indicate sentence-level correlations in contrast to a token-level analysis. *Left:* Sentiment Reading on the SST dataset. *Right:* Relation Extraction on Wikipedia. Standard deviations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

attention matrices as a graph, where tokens are represented as nodes and attention scores as edges between consecutive layers. The edge values define the maximal flow possible between a pair of nodes. Flow between edges is thus (i) limited to the maximal attention between any two consecutive layers for this token and (ii) conserved such that the sum of incoming flow must be equal to the sum of outgoing flow. We denote the attention flow propagated back from layer L as *flow* L.

Shallow Models We ground our analysis on Transformers by comparing them to relatively shallow models that were trained from-scratch and evaluate how well they coincide with human fixation. We train a standard **CNN** (Kim, 2014) network with multiple filter sizes on pre-trained GloVe embeddings (Pennington et al., 2014). Importance scores over tokens are extracted using Layerwise Relevance Propagation (LRP) (Arras et al., 2016, 2017) which has been demonstrated to produce robust explanations by iterating over layers and redistributing relevance from outer layers towards the input (Bach et al., 2015; Samek et al., 2021). In parallel, we use a shallow multi-head **self-attention** network (Lin et al., 2017) on GloVe vectors with a linear read-out layer for which we compute token relevance scores using LRP.

E-Z Reader As a cognitive model for human reading, we compute task-neutral fixation times using the E-Z Reader (Reichle et al., 1998) model. The E-Z Reader is a multi-stage, hybrid model, which relies on an n-gram model and several heuristics, based, for example, on theoretical assumptions about the role of predictability and average saccade length. Additionally, we compare to a frequency baseline using word statistics of the **BNC** (British National Corpus, Kilgarriff (1995))[1] as proposed by Barrett et al. (2018).

3.3 Optimization

For training models on the different tasks we remove all sentences that overlap between ZuCo and the original SST and Wikipedia datasets. Models are then trained on the remaining train-split data until early stopping is reached and we report results over five runs. We provide further details on the optimization and model task performance in Appendix A.

3.4 Metric

To compare models with human attention, we compute Spearman correlation between human and model-based importance vectors after concatenation of individual sentences as well as on a token-level, see Hollenstein and Beinborn (2021). This enables us to distinguish unrelated effects caused by varying sentence length from token-level importance. As described before, we extract human attention from gaze (ZuCo), simulated gaze (E-Z Reader), raw attentions (BERT, RoBERTa, T5), relevance scores (CNN, self-attention) and inverse token probability scores (BNC).[2] We use ZuCo to-

[1] We compute the negative log-transformed probability of each lower-cased token corresponding to an inverse relation between word-frequency and human gaze duration (Rayner and Duffy, 1986)

[2] First and last token bins from each sentence are ignored to avoid the influence of sentence border effects in Transformers (Clark et al., 2019) and for which the E-Z Reader does not compute fixations.

kens to align sentences across tokenizers and apply max-pooling of scores when bins are merged.

3.5 Main result

To evaluate how well model and human attention patterns for sentiment reading and relation extraction align, we compute pair-wise correlation scores as displayed in Figure 1. Reported correlations are statistically significant with $p < 0.01$ if not indicated otherwise (ns: not significant). After ranking based on the correlations on sentence-level, we observe clear differences between sentiment reading on SST and relation extraction on Wikipedia for the different models. For sentiment reading, the E-Z Reader and BNC show the highest correlations followed by the Transformer attention flow values (the ranking between E-Z/BNC and Transformer flows is significant at $p < 0.05$). For relation extraction, we see the highest correlation for BERT-base attention flows (with and without fine-tuning) and BERT-large followed by the E-Z Reader (ranking is significant at $p < 0.05$). On the lower end, computing means over BERT attentions across the last layer shows weak to no correlations for both tasks.[3] The shallow architectures result in low to moderate correlations with a distinctive gap to attention flow. Focusing on flow values for Transformers, BNC and E-Z Reader, correlations are stable across word and sentence length. Correlations grouped by sentence length shows stable values around 0.6 (SST) and $0.4 - 0.6$ (Wikipedia) except for shorter sentences where correlations fluctuate. To check the linear relationship between human and model attention patterns we additionally compute token- and sentence-level Pearson correlations which can be found in Appendix B. Results confirm that Spearman and Pearson correlation coefficients as well as rankings hardly differ - which suggests a linear relationship - and that correlation strength is in line with Hollenstein and Beinborn (2021).

4 Analyses

In addition to our main result – that pre-trained language models *are* competitive to heuristic cognitive models in predicting human eye fixations during reading – we present a detailed analysis, investigating what our main results depend on, where pre-trained language models improve on cognitive models, and where they are still challenged.

Fine-tuning BERT does not change correlations to human attention We find that fine-tuning base and large BERT models on either task does not significantly change correlations and are of similar strength to untuned models. This observation can be embedded into findings that Transformers are equipped with overcomplete sets of attention functions that hardly change until the later layers, if at all, during fine-tuning and that this change is also dependent on the tuning task itself (Kovaleva et al., 2019; Zhao and Bethard, 2020). In addition, we observe that Transformer flows propagated back from early, medium and final layers do not considerably change correlations to human attention. This can be explained by attention flow filtering the path of minimal value at each layer as discussed in Abnar and Zuidema (2020).

Attention flow is important The correlation analysis emphasizes that we need to capture the layered propagation structure in Transformer models, e.g., by using attention flow, in order to extract importance scores that are competitive with cognitive models. Interestingly, selecting the highest correlating head for the last attention layer produces generally weaker correlation than attention flows.[3] This offers additional evidence that raw attention weights do not reliably correspond to token relevance (Serrano and Smith, 2019; Abnar and Zuidema, 2020) and, thus, are of limited use to compare task attention to human gaze.

Differences between language models BERT, RoBERTa and T5 are large-scale pretrained language models based on Transformers, but they also differ in various ways. One key difference is that BERT and RoBERTa use absolute position encodings, while T5 uses relative encodings. BERT and RoBERTa differ in that (i) BERT has a next-sentence-prediction auxiliary objective; (ii) RoBERTa and T5 were trained on more data; (iii) RoBERTa uses dynamic masking and trains with larger mini-batches and learning rates, while T5 uses multi-word masking; (iv) RoBERTa uses byte pair encoding for subword segmentation. We leave it as an open question whether the superior attention flows of BERT, compared to RoBERTa and T5, has to do with training data, next sentence prediction, or fortunate hyper-parameter settings, but note that BERT is also known to have

[3] We have experimented with oracle analyses selecting the maximally correlating attention head in the last layer for each sentence and find that correlations are generally weaker than with attention flow.

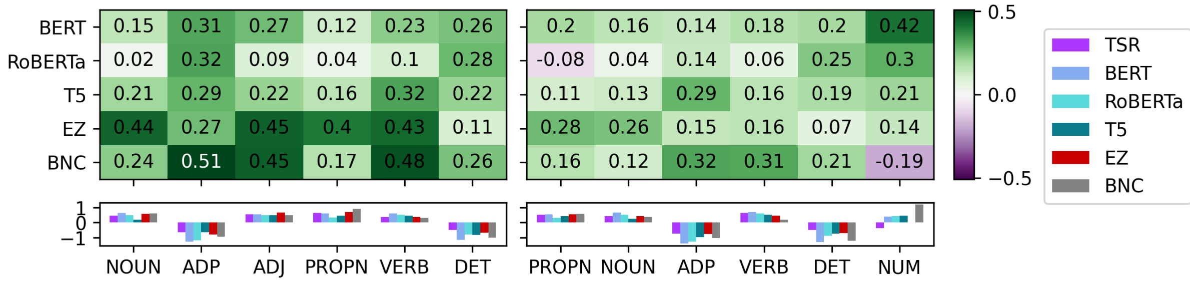

Figure 2: *Upper:* Correlations between human fixation and different models for SST *(left)* and Relation Extraction *(right)* for the six most common POS tags. *Lower:* Average attention value after standardization (mean=0, std=1) for respective POS tag and model.

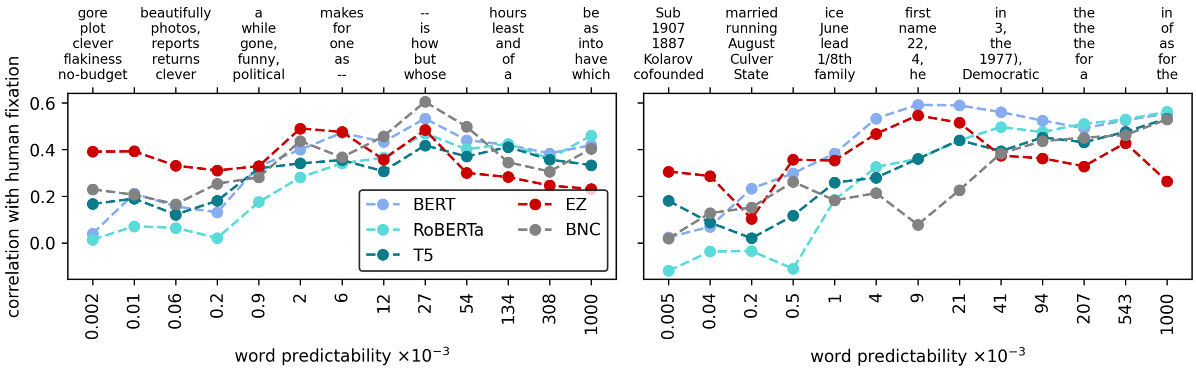

Figure 3: Correlation between human fixations and different models for SST *(left)* and Wikipedia *(right)* with respect to word predictability in equally sized bins. Word predictability scores, were calculated with a 5-gram Kneser-Ney language model. Respective bin limits are given on the x-axis. Samples for every other bin are displayed on the upper x-axis.

higher alignment with human-generated explanations than other large-scale pre-trained language models (Prasad et al., 2021).

E-Z Reader is less sensitive to hard-to-predict words and POS 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 computed with a 5-gram Kneser-Ney language model (Kneser and Ney, 1995; Chelba et al., 2013). Figure 3 shows the results on SST and Wikipedia for equally sized bins of word predictability scores. We can see that the Transformer models correlate better for more predictable words on both datasets whereas the E-Z Reader is less influenced by word predictability and already shows medium correlation on the most hard-to-predict words ($0.3 - 0.4$ for both, SST and Wikipedia). In fact, on SST, Transformers only pass the E-Z Reader on the most predictable tokens (word predictability > 0.03).

We also compare correlations to human fixations based on the top-6 (most tokens) Part-of-speech (POS) tags. On SST, correlations with E-Z Reader are very consistent across POS tags whereas attention flow shows weak correlations on proper nouns (0.12), nouns (0.16) and verbs (0.16) as presented in Figure 2. 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.

Input reduction When comparing machines to humans we typically regard the psychophysical data as the gold standard. We will now take the model perspective and test fidelity of both human and model attention patterns in task-tuned models. By this we aim to test how effective the exact token ranking based on attention scores is at producing the correct output probability. We perform such an input reduction analysis (Feng et al., 2018) using fine-tuned BERT models for both sentiment classification and relation extraction as the reference model and present results in Figure 4. In

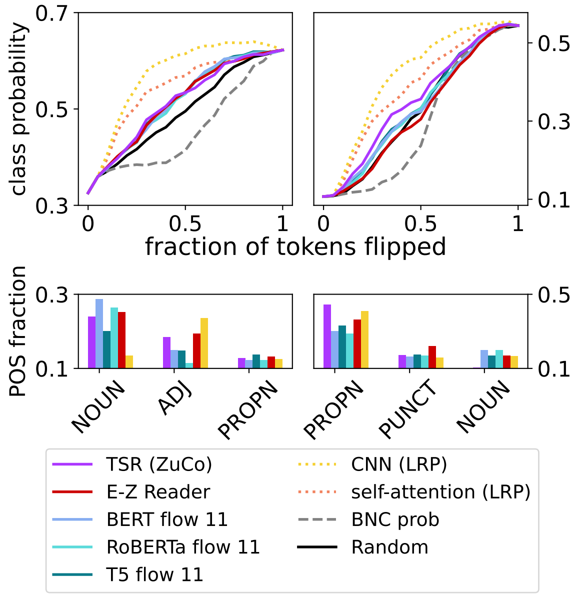

Figure 4: Results of our reduction analysis where most important tokens are selected and fed into fine-tuned BERT models for sentiment classification *(left)* and relation extraction *(right)*. *Upper:* we gradually measure output probability for the true label. Higher area under the curve reflects a 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.

our analysis, we observe - as to be expected - that adding tokens according to token probability (BNC prob) performs even worse than randomly adding tokens. From-scratch trained models (CNN and self-attention) are most effective in selecting task-relevant tokens, and even more so than using any Transformer attention flow. Adding tokens based on human attention is as effective for the sentiment task as the E-Z Reader. Interestingly, for the relation extraction task, human attention vectors provide the most effective flipping order after the relevance-based shallow methods. All Transformer-based flows perform comparably in both tasks. To better understand what drives these effects we extract the fraction of POS tags for the first added token (see Figure 4 and full results in the Appendix Figure 5). For sentiment reading, the flipping according to CNN relevances puts more emphasis on adjectives (ADJ) whereas the other methods tend to flip nouns (NOUN) first. Across the Transformer models RoBERTa relies much less on adjectives than any other model. In the relation extraction task, we observe that proper nouns (PROPN) are dominant (and adjectives play almost no role) in all model systems which highlights the role of task nature on the importance assignment. In addition,

	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 1: 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.

we see that the E-Z Reader overestimates the importance of punctuation, whereas proper nouns are least dominant in comparison to the other models.

Entropy levels of Transformer flow is similar to those in human attention Averaged sentence-level entropy values on both datasets reveal that BERT, RoBERTa and T5 attention flow, the E-Z Reader and BNC obtain similar levels of sparsity as human attention around 3.4-3.6 bits as summarized in Table 1. Entropies are lower for the shallow networks with self-attention (LRP) at 1.8-2.2 bits and CNN (LRP) at around 2.9 bits. This difference in sparsity levels might explain the advantage of CNN and shallow self-attention in the input reduction analysis: Early addition of few but very relevant words has a strong effect on the model's decision compared to less sparse scoring as, e.g. in Transformers. The shallow models were also trained from-scratch for the respective tasks whereas all other models (including human attention) are heavily influenced by a more general modeling of language which could explain attention to be distributed more broadly over all tokens.

	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

Table 2: Correlations between human fixations and models on 48 duplicates appearing in the ZuCo dataset for both natural reading (NR) and relation extraction (task-specific reading - TSR).

Natural reading versus task-specific reading A unique feature of the ZuCo dataset is that it contains a subset of sentences that were presented to participants both in a task-specific (relation extraction) and a natural reading setting. This allows for a direct comparison of how correlation strength is influenced by the task. In Table 2 correlations of human gaze-based attention with model attentions are shown. The highest correlation can be observed when comparing human attention for task-specific and natural reading (0.72). The remaining model correlations correspond to the ranking and correlation strength observed in the main result (see Figure 1). We observe lower correlation scores for the task-specific reading as compared to normal reading among attention flow, the E-Z Reader and BNC. This suggests that these models capture the statistics of natural reading - as is expected for a cognitive model designed to the natural reading paradigm - and that task-related changes in human fixation patterns are not reflected in Transformer attention flows. Interestingly, averaged last layer attention heads show a reverse effect (but at much weaker correlation strength). This might suggest that pre-training in Transformer models induces specificity of later layer attention heads to task-solving instead of general natural reading patterns.

5 Related Work

Saliency modeling Early computational models of visual attention have used bottom-up approaches to model the neural circuitry representing pre-attentive selection processes from visual input (Koch and Ullman, 1985) and later the central idea of a saliency map was introduced (Niebur and Koch, 1996). A central hypothesis studying eye movements under task conditions is known as Yarbus theorem stating that a task can be directly decoded from fixation patterns (Yarbus, 1967) which has found varying support (Greene et al., 2012; Henderson et al., 2013; Borji and Itti, 2014).

More recently, extracting features from deep pre-trained filters in combination with readout networks has boosted performance on the saliency task (Kümmerer et al., 2016). This progress has enabled modeling of more complex gaze patterns, e.g. vision-language tasks such as image captioning (Sugano and Bulling, 2016), visual question answering (Das et al., 2016) or text-guided object detection (Vasudevan et al., 2018).

Predicting text gaze patterns has been studied extensively, often in the context of probabilistic (Feng, 2006; Hara et al., 2012; Matthies and Søgaard, 2013; Hahn and Keller, 2016) or token transition models (Nilsson and Nivre, 2009; Haji-Abolhassani and Clark, 2014; Coutrot et al., 2017). More recently deep language features have been used as feature extractors in modeling text saliency (Sood et al., 2020a; Hollenstein et al., 2021) opening the question of their cognitive plausibility.

Eye-tracking signals for NLP Augmenting machine learning models using human gaze information has been shown to improve performance for a number of different settings: Human attention patterns as regularization during model training have resulted in comparable or improved task performance in tagging part-of-speech (Barrett and Søgaard, 2015a,b; Barrett et al., 2018), sentence compression (Klerke et al., 2016), detecting sentiment (Mishra et al., 2016, 2017) or reading comprehension (Malmaud et al., 2020). In these works, general free-viewing gaze data is used without consideration of the specific training task which opens the question of task-modulation in human reading.

From natural to task-specific reading Recent work on reading often analyses eye-tracking data in combination with neuroimaging techniques such as EEG (Wenzel et al., 2017) and f-MRI (Hillen et al., 2013; Choi et al., 2014). Research questions thereby focus either on detecting relevant parts in text (Loboda et al., 2011; Wenzel et al., 2017) or the difference between natural and pseudo-reading, i.e., text without syntax/semantics (Hillen et al., 2013) or pseudo-words (Choi et al., 2014). To the best of our knowledge there has not been any work on comparing fixations between natural reading and task-specific reading on classical NLP tasks such as relation extraction or sentiment classification.

6 Discussion and Conclusion

In this paper, we have compared attention and relevance mechanisms of a wide range of models to human gaze patterns when solving sentiment classification on SST movie reviews and relation extraction on Wikipedia articles. We generally found that Transformer architectures are competitive with the E-Z Reader, but only when computing attention flow scores. We generally saw weaker correlations for relation extraction on Wikpedia, presumably due to simpler sentence structures and the occurrence of polarity words. In the following, we discuss implications of our findings on NLP and Cognitive Science in more detail.

Lessons for NLP One implication of the above for NLP follows from the importance of attention flow in our experiments: Using human gaze to regularize or supervise attention weights has proven effective in previous work (§5), but we observed that correlations with task-specific human attention increase significantly by using layer-dependent attention flow compared to using raw attention weights. This insight motivates going beyond regularizing raw attention weights or directly injecting human attention vectors during training, to instead optimize for correlation between attention flow and human attention. Jointly modeling language and human gaze has recently shown to yield competitive performance on paraphrase generation and sentence compression while resulting in more task-specific attention heads (Sood et al., 2020b). For this study natural gaze patterns were also simulated using the E-Z Reader.

Another potential implication concerns interpretability. It remains an open problem how best to interpret self-attention modules (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), and whether they provide meaningful explanations for model predictions. Including gradient information to explain Transformers has recently been considered to improve their interpretability (Chefer et al., 2021b,a; Ali et al., 2022). A successful explanation of a machine learning model should be faithful, human-interpretable and practical to apply (Samek et al., 2021). Faithfulness and practicality is often evaluated using automated procedures such as input reduction experiments or measuring time and model complexity. By contrast, judging human-interpretability typically requires costly experiments in well-controlled settings and obtaining human gold-standards for interpretability remain difficult (Miller, 2019; Schmidt and Bießmann, 2019). Using gaze data to evaluate the faithfulness and trustworthiness of machine learning models is a promising approach to increase model transparency.

Lessons for Cognitive Science Attention flow in Transformers, especially for BERT models, correlates surprisingly well with human task-specific reading, but what does this tell us about the shortcomings of our cognitive models? We know that word frequency and semantic relationships between words influence word fixation times (Rayner, 1998). In our experiments, we see relatively high correlation between human fixations and the inverse word probability baseline which raises the question to what extent reading gaze is driven by low-level patterns such as word frequency or syntactic structure in contrast to more high-level semantic context or wrap-up effects.

In computer vision, cognitively inspired bottom-up models, e.g., using intensity and contrast features, are able to explain at most half of the gaze fixation information in comparison to the human gold standard (Kümmerer et al., 2017). The robustness of the E-Z Reader on movie reviews is likely due to its explicit modeling of low-level properties such as word frequency or sentence length. BERT was recently shown to be primarily modeling higher-order word co-occurrence statistics (Sinha et al., 2021). We argue that while Transformers are limited, e.g., in not capturing the dependency of human gaze on word length (Kliegl et al., 2004), cognitive models seem to underestimate the role of word co-occurrence statistics.

During reading, humans are faced with a trade-off between the precision of reading comprehension and reading speed, by avoiding unnecessary fixations (Hahn and Keller, 2016). This trade-off is related to the input reduction experiments performed in Section 4. Here, we observe that shallow methods score well at being sparse and effective in changing model output towards the correct class, but produce only weak correlation to human reading patterns when compared to layered language models. In comparison, extracted attention flow from pre-trained Transformer models correlates much better with human attention, but offers less sparse token attention. In other words, our results show that task-specific reading is sub-optimal relative to solving tasks and heavily regularized by natural reading patterns (see also our comparison of task-specific and natural reading in Section 4).

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 pretrained 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.

Acknowledgements

This work was partially funded by the German Ministry for Education and Research as BIFOLD – Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A), as well as by the Platform Intelligence in News project, which is supported by Innovation Fund Denmark via the Grand Solutions program. We thank Mostafa Abdou for fruitful discussions and Heather Lent, Miryam de Lhoneux and Vinit Ravishankar for proof-reading and valuable inputs on the manuscript.

References

Mostafa Abdou, Artur Kulmizev, Felix Hill, Daniel M. Low, and Anders Søgaard. 2019. Higher-order comparisons of sentence encoder representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5838–5845, Hong Kong, China. Association for Computational Linguistics.

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. 2022. XAI for transformers: Better explanations through conservative propagation. *CoRR*, abs/2202.07304.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. Explaining predictions of non-linear classifiers in NLP. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 1–7, Berlin, Germany. Association for Computational Linguistics.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):1–23.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, Brussels, Belgium. Association for Computational Linguistics.

Maria Barrett and Anders Søgaard. 2015a. Reading behavior predicts syntactic categories. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 345–349, Beijing, China. Association for Computational Linguistics.

Maria Barrett and Anders Søgaard. 2015b. Using reading behavior to predict grammatical functions. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, pages 1–5, Lisbon, Portugal. Association for Computational Linguistics.

Ali Borji and Laurent Itti. 2014. Defending yarbus: eye movements reveal observers' task. *Journal of vision*, 14(3):29.

Hila Chefer, Shir Gur, and Lior Wolf. 2021a. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406.

Hila Chefer, Shir Gur, and Lior Wolf. 2021b. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. Technical report, Google.

Wonil Choi, Rutvik H Desai, and John M Henderson. 2014. The neural substrates of natural reading: a comparison of normal and nonword text using eye-tracking and fmri. *Frontiers in human neuroscience*, 8:1024.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

Kenneth Church and Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:10.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP:*

Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Antoine Coutrot, Janet H. Hsiao, and Antoni B. Chan. 2017. Scanpath modeling and classification with hidden markov models.

Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 296–303.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gary Feng. 2006. Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research*, 7(1):70–95. Models of Eye-Movement Control in Reading.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. 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. Association for Computational Linguistics.

Michelle Greene, Tommy Liu, and Jeremy Wolfe. 2012. Reconsidering yarbus: A failure to predict observers' task from eye movement patterns. *Vision research*, 62:1–8.

Fritz Günther, Luca Rinaldi, and Marco Marelli. 2019. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, 14(6):1006–1033. PMID: 31505121.

Michael Hahn and Frank Keller. 2016. Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 85–95, Austin, Texas. Association for Computational Linguistics.

Amin Haji-Abolhassani and James J. Clark. 2014. An inverse yarbus process: Predicting observers' task from eye movement patterns. *Vision Research*, 103:127–142.

Tadayoshi Hara, Daichi Mochihashi, Yoshinobu Kano, and Akiko Aizawa. 2012. Predicting word fixations in text with a CRF model for capturing general reading strategies among readers. In *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*, pages 55–70, Mumbai, India. The COLING 2012 Organizing Committee.

John Henderson, Svetlana Shinkareva, Jing Wang, Steven Luke, and Jenn Olejarczyk. 2013. Predicting cognitive state from eye movements. *PloS one*, 8:e64937.

Rebekka Hillen, Thomas Günther, Claudia Kohlen, Cornelia Eckers, Muna van Ermingen-Marbach, Katharina Sass, Wolfgang Scharke, Josefine Vollmar, Ralph Radach, and Stefan Heim. 2013. Identifying brain systems for gaze orienting during reading: fmri investigation of the landolt paradigm. *Frontiers in human neuroscience*, 7:384.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. *arXiv preprint arXiv:2104.05433*.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Kilgarriff. 1995. BNC database and word frequency lists. *Accessed: 07/2020*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533, San Diego, California. Association for Computational Linguistics.

Sigrid Klerke and Barbara Plank. 2019. At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Christof Koch and Shimon Ullman. 1985. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, 4:219–227.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv:1610.01563 [cs, q-bio, stat]*. ArXiv: 1610.01563.

Matthias Kümmerer, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge. 2017. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808.

Yann Lecun and Yoshua Bengio. 1995. *Convolutional Networks for Images, Speech and Time Series*, pages 255–258. The MIT Press.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomasz D Loboda, Peter Brusilovsky, and Jöerg Brunstein. 2011. Inferring word relevance from eye-movements of readers. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 175–184.

Jonathan Malmaud, Roger Levy, and Yevgeni Berzak. 2020. Bridging information-seeking human gaze and machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 142–152, Online. Association for Computational Linguistics.

Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting : a review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Franz Matthies and Anders Søgaard. 2013. With blinkers on: Robust prediction of eye movements across readers. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 803–807, Seattle, Washington, USA. Association for Computational Linguistics.

Sara V. Milledge and Hazel I. Blythe. 2019. The changing role of phonology in reading development. *Vision*, 3(2).

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–387, Vancouver, Canada. Association for Computational Linguistics.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Leveraging cognitive features for sentiment analysis. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 156–166, Berlin, Germany. Association for Computational Linguistics.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham.

E. Niebur and C. Koch. 1996. Control of selective visual attention: Modeling the "where" pathway. In D. S Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 802–808. MIT Press, Cambridge, MA.

Mattias Nilsson and Joakim Nivre. 2009. Learning where to look: Modeling eye movements in reading. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning, CoNLL 2009, Boulder, Colorado, USA, June 4-5, 2009*, pages 93–101. ACL.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. 2021. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 1–14, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Memory amp; cognition*, 14(3):191—201.

Keith Rayner and Erik D. Reichle. 2010. Models of the reading process. *WIREs Cognitive Science*, 1(6):787–799.

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.

Erik D. Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The e-z reader model of eye-movement control in reading: comparisons to other models. *The Behavioral and brain sciences*, 26(4):445–476.

Timothy Rogers and Michael Wolmetz. 2016. Conceptual knowledge representation: A cross-section of current research. *Cognitive Neuropsychology*, 33:1–9.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.

Philipp Schmidt and Felix Bießmann. 2019. Quantifying interpretability and trust in machine learning systems. *CoRR*, abs/1901.08558.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Anders Søgaard. 2016. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, Berlin, Germany. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Mueller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 6327–6341. Curran Associates, Inc.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002*, pages 901–904.

Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *CoRR*, abs/1608.05203.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2018. Object referring in videos with language and human gaze. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4129–4138. IEEE Computer Society.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Markus Andreas Wenzel, Mihail Bogojeski, and Benjamin Blankertz. 2017. Real-time inference of word relevance from electroencephalogram and eye gaze. *Journal of neural engineering*, 14(5):056007.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Alfred L. Yarbus. 1967. *Eye Movements and Vision*. Plenum. New York.

Yingyi Zhang and Chengzhi Zhang. 2019. Using human attention to extract keyphrase from microblog post. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5867–5872, Florence, Italy. Association for Computational Linguistics.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.

A Model and Optimization Details

In the following we present details for all modes and describe the training details used for task-tunning. Model performance over five runs is reported in Table 3.

A.1 CNN

The CNN models use 300-dimensional pre-trained GloVe_840B (Pennington et al., 2014) embeddings. Input sentences are tokenized using the SpaCy tokenizer (Honnibal et al., 2020). We use 150 convolutional filters of filter sizes $s = [3, 4, 5]$ with ReLu activation, followed by a max-pooling-layer and apply dropout of $p = 0.5$ of the linear classification layer during training. For training we use a batchsize of $bs = 50$ and train all model parameters using the Adam optimizer with a learning rate of $lr = 1e-4$ for a maximum number of $T = 20$ epochs. For all model trainings, we apply early stopping to avoid overfitting during training and stop optimization as soon as the validation loss begins to increase. To compute LRP relevances we use the general formulation of LRP propagation rules with $\gamma = 0.$ for the linear readout layers (Montavon et al., 2019). We take absolute values over resulting relevance scores since we find they correlate best with human attention in comparison to raw and rectified processing. For propagation through the max-pooling layer we apply the winner-take-all principle and for convolutional layers we use the LRP-γ redistribution rule and select $\gamma = 0.5$ after a search over $\gamma = [0., 0.25, 0.5, 0.75, 1.0]$ resulting in largest correlations to human attention.

A.2 Self-Attention model

For the multi-head self-attention model again use 300-dimensional pre-trained GloVe_840B embeddings and tokenized via SpaCy. The architecture consists of a set of $k = 3$ self-attention heads for the SR task and $k = 8$ for REL. The resulting sentence representation is then fed into a linear classification readout layer with $\gamma = 0.$ and which we also use for the propagation to input embeddings. During optimization we use $lr = 1e-4$, $bs = 50$ and $T = 50$.

A.3 Transformer Models

We use standard BERT-base/large-uncased architectures and tokenizers as provided by the huggingface library (Wolf et al., 2020). For BERT-base fine-tuning we use $lr = 1e-5$ for REL and $lr = 1e-6$ for SR, $bs = 32$ and $T = 50$ for both tasks. For BERT-large we use $lr = 1e-5$ for REL and $lr = 5e-7$ for SR, $bs = 16$ and $T = 50$. For RoBERTa and T5 we use the RoBERTa-base and T5-base checkpoints and respective tokenizers.

A.4 E-Z Reader

We use version 10.2 of the E-Z Reader with default parameters and 1000 repetitions. Cloze scores, i.e. word predictability scores, were therefore computed using a 5-gram **Kneser-Ney** language model (Kneser and Ney, 1995) as provided by the SRI Language Modeling Toolkit (Stolcke, 2002) and

	Acc (SR)	F1 (SR)	Acc (REL)	F1 (REL)
self-attention	69.0 ± 0.2	64.5 ± 2.2	67.5 ± 1.3	55.5 ± 2.0
CNN	71.3 ± 0.2	69.8 ± 1.7	74.0 ± 1.9	68.7 ± 4.8
BERT-base	76.0 ± 0.1	67.0 ± 3.0	78.3 ± 1.5	72.7 ± 3.3
BERT-large	76.4 ± 0.1	63.8 ± 1.3	78.9 ± 2.3	71.0 ± 2.7

Table 3: Accuracy and F1 scores after fine-tuning on the respective task dataset over five runs: sentiment reading on SST (SR) and relation extraction on Wikipedia (REL). Samples that overlap with the ZuCo dataset were filtered out.

trained on the 1 billion token dataset (Chelba et al., 2013). Resulting perplexity on the held-out test set was $ppl = 81.9$. Then, word-based total fixation times are computed from the E-Z Readers trace files and averaged over all subjects.

B Spearman versus Pearson correlation on sentence and token level

In addition to Spearman correlation over all tokens, we also report Pearson correlation coefficients on a sentence and token-level. Results are displayed in Table 4. Compared to Spearman correlation on all tokens, the ranking does hardly change for Pearson or sentence-level correlations. Absolute correlation coefficients are higher for Spearman compared to Pearson and also are slightly higher on the sentence-level as compared to the token-level analysis. Biggest changes occur in a drop for BNC when Spearman correlation is calculated on all tokens for relation extraction and an increase for self-attention (LRP) in sentiment reading. We hypothesize that both effects can be traced back to the level of sparsity and the corresponding ranking for Spearman correlations. In our entropy analysis we found that, i.e. self-attention shows a sparser representation which was likely caused by the overconfidence of the model, and which could explain the higher rank-based correlation.

C Input reduction - POS tag analysis

Figure 5 shows the full distribution of POS tags of the first tokens flipped. This extends Figure 4 where we only show the first 3 POS tags.

D Entropy analysis

We compute entropy values for different attention and relevance scores in both task settings. To compensate for different sentence lengths we perform a stratified analysis such that every sentence length occurs equally often in both tasks. Sentence lengths

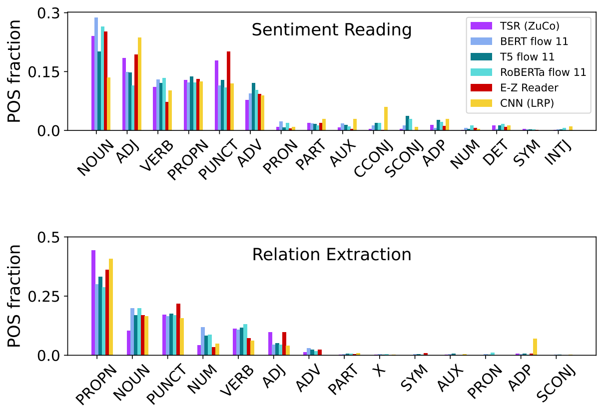

Figure 5: Full distribution of POS tags of most important first flip tokens for the task of sentiment reading (top) and relation extraction (bottom).

which merely occur in one of the two tasks, are excluded from the sampling. Maximum entropy is reached for uniformly distributed token scores.

	SR				TSR			
	tok		sent		tok		sent	
	pearson	spearman	pearson	spearman	pearson	spearman	pearson	spearman
BNC inv prob	0.57	0.66	0.62	0.64	0.34	0.41	0.45	0.46
CNN (LRP)	0.17	0.27	0.27	0.26	0.13	0.14	0.21	0.18
self-attention	0.36	0.48	0.43	0.54	0.27	0.49	0.44	0.61
self-attention (LRP)	0.07	0.39	0.34	0.43	0.09	0.31	0.28	0.36
BERT flow 0	0.52	0.62	0.61	0.63	0.47	0.55	0.55	0.60
BERT flow 5	0.53	0.61	0.60	0.61	0.49	0.53	0.57	0.57
BERT flow 11	0.54	0.62	0.62	0.63	0.51	0.56	0.60	0.61
fine-BERT flow 0	0.52	0.62	0.61	0.63	0.47	0.55	0.55	0.60
fine-BERT flow 5	0.53	0.61	0.59	0.61	0.50	0.54	0.59	0.59
fine-BERT flow 11	0.54	0.62	0.62	0.63	0.51	0.56	0.60	0.60
BERT-large flow 0	0.51	0.61	0.62	0.63	0.47	0.54	0.57	0.60
BERT-large flow 11	0.55	0.63	0.62	0.62	0.50	0.55	0.57	0.57
BERT-large flow 23	0.55	0.63	0.62	0.62	0.50	0.55	0.57	0.57
fine-BERT-large flow 0	0.51	0.61	0.62	0.63	0.47	0.54	0.57	0.60
fine-BERT-large flow 11	0.55	0.63	0.62	0.62	0.50	0.55	0.57	0.57
fine-BERT-large flow 23	0.55	0.63	0.62	0.62	0.50	0.55	0.57	0.57
RoBERTa flow 0	0.44	0.54	0.52	0.55	0.35	0.43	0.42	0.47
RoBERTa flow 5	0.32	0.42	0.45	0.46	0.26	0.33	0.36	0.40
RoBERTa flow 11	0.44	0.51	0.51	0.52	0.37	0.41	0.45	0.46
T5 flow 0	0.44	0.53	0.51	0.54	0.37	0.44	0.47	0.50
T5 flow 5	0.43	0.50	0.49	0.49	0.35	0.40	0.44	0.43
T5 flow 11	0.44	0.51	0.51	0.53	0.37	0.42	0.46	0.46
BERT mean	0.04	0.14	0.10	0.11	-0.03	0.11	0.02	0.09
fine-BERT mean	0.03	0.09	0.05	0.03	-0.03	0.10	0.02	0.08
BERT-large mean	-0.01	0.20	0.10	0.28	-0.03	0.14	-0.01	0.14
fine-BERT-large mean	-0.02	0.11	0.04	0.17	-0.09	-0.05	-0.12	-0.06
RoBERTa mean	0.22	0.22	0.26	0.21	0.08	0.10	0.14	0.10
T5 mean	-0.00	0.06	-0.00	0.07	-0.02	0.10	0.02	0.19
E-Z Reader	0.64	0.65	0.69	0.67	0.46	0.51	0.56	0.56

Table 4: Full correlation analysis for sentiment reading *(left)* and relation extraction *(right)*. We show Spearman and Pearson correlation coefficients between human fixations and models. Correlation coefficients were calculated per sentence and averaged (sen) or after concatenation of all sentences (tok)