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

Eberle, Oliver; Brandl, Stephanie; Pilot, Jonas; Søgaard, Anders

Published in:
ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)

DOI:
10.18653/v1/2022.acl-long.296

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Do Transformer Models Show Similar Attention Patterns to Task-Specific Human Gaze?

Oliver Eberle∗,1, Stephanie Brandl∗,1,2, Jonas Pilot1, Anders Søgaard2
1Machine Learning Group, TU Berlin, Germany 2University 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), 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.

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.
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; Mander 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
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

\(^3\)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.

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
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.

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.
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, 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Entropy</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>3.47</td>
<td>0.15</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>T5</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>Fine-BERT</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>E-Z Reader</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>BNC</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>CNN</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>Self-attention (LRP)</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>BNC prob</td>
<td>3.44</td>
<td>0.15</td>
</tr>
<tr>
<td>Random</td>
<td>3.44</td>
<td>0.15</td>
</tr>
</tbody>
</table>

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 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 (Kümerer 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; Matthis 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 Wikipedia, 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-interpretable 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 models is a promising approach to increase model transparency.
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.

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 Ravisankar for proof-reading and valuable inputs on the manuscript.

References


Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all...


A Model and Optimization Details

In the following we present details for all modes and describe the training details used for task-tuning. 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-\( \gamma \) redistribution rule and select \( \gamma = 0.5 \) after a search over \( \gamma = [0.0, 0.25, 0.5, 0.75, 1.0] \) 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
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.

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

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 over-confidence 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 which merely occur in one of the two tasks, are excluded from the sampling. Maximum entropy is reached for uniformly distributed token scores.
<table>
<thead>
<tr>
<th></th>
<th>tok</th>
<th>sent</th>
<th>tok</th>
<th>sent</th>
<th>tok</th>
<th>sent</th>
<th>tok</th>
<th>sent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pearson</td>
<td>spearman</td>
<td>pearson</td>
<td>spearman</td>
<td>pearson</td>
<td>spearman</td>
<td>pearson</td>
<td>spearman</td>
</tr>
<tr>
<td>BNC inv prob</td>
<td>0.57</td>
<td>0.66</td>
<td>0.62</td>
<td>0.64</td>
<td>0.34</td>
<td>0.41</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>CNN (LRP)</td>
<td>0.17</td>
<td>0.27</td>
<td>0.27</td>
<td>0.26</td>
<td>0.13</td>
<td>0.14</td>
<td>0.21</td>
<td>0.18</td>
</tr>
<tr>
<td>self-attention</td>
<td>0.36</td>
<td>0.48</td>
<td>0.43</td>
<td>0.54</td>
<td>0.27</td>
<td>0.49</td>
<td>0.44</td>
<td>0.61</td>
</tr>
<tr>
<td>self-attention (LRP)</td>
<td>0.07</td>
<td>0.39</td>
<td>0.34</td>
<td>0.43</td>
<td>0.09</td>
<td>0.31</td>
<td>0.28</td>
<td>0.36</td>
</tr>
<tr>
<td>BERT flow 0</td>
<td>0.52</td>
<td>0.62</td>
<td>0.61</td>
<td>0.63</td>
<td>0.47</td>
<td>0.55</td>
<td>0.55</td>
<td>0.60</td>
</tr>
<tr>
<td>BERT flow 5</td>
<td>0.53</td>
<td>0.61</td>
<td>0.60</td>
<td>0.61</td>
<td>0.49</td>
<td>0.53</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>BERT flow 11</td>
<td>0.54</td>
<td>0.62</td>
<td>0.62</td>
<td>0.63</td>
<td>0.51</td>
<td>0.56</td>
<td>0.60</td>
<td>0.61</td>
</tr>
<tr>
<td>fine-BERT flow 0</td>
<td>0.52</td>
<td>0.62</td>
<td>0.61</td>
<td>0.63</td>
<td>0.47</td>
<td>0.55</td>
<td>0.55</td>
<td>0.60</td>
</tr>
<tr>
<td>fine-BERT flow 5</td>
<td>0.53</td>
<td>0.61</td>
<td>0.59</td>
<td>0.61</td>
<td>0.50</td>
<td>0.54</td>
<td>0.59</td>
<td>0.59</td>
</tr>
<tr>
<td>fine-BERT flow 11</td>
<td>0.54</td>
<td>0.62</td>
<td>0.62</td>
<td>0.63</td>
<td>0.51</td>
<td>0.56</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>BERT-large flow 0</td>
<td>0.51</td>
<td>0.61</td>
<td>0.62</td>
<td>0.63</td>
<td>0.47</td>
<td>0.54</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>BERT-large flow 11</td>
<td>0.55</td>
<td>0.63</td>
<td>0.62</td>
<td>0.62</td>
<td>0.50</td>
<td>0.55</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>BERT-large flow 23</td>
<td>0.55</td>
<td>0.63</td>
<td>0.62</td>
<td>0.62</td>
<td>0.50</td>
<td>0.55</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>fine-BERT-large flow 0</td>
<td>0.51</td>
<td>0.61</td>
<td>0.62</td>
<td>0.63</td>
<td>0.47</td>
<td>0.54</td>
<td>0.57</td>
<td>0.60</td>
</tr>
<tr>
<td>fine-BERT-large flow 11</td>
<td>0.55</td>
<td>0.63</td>
<td>0.62</td>
<td>0.62</td>
<td>0.50</td>
<td>0.55</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>fine-BERT-large flow 23</td>
<td>0.55</td>
<td>0.63</td>
<td>0.62</td>
<td>0.62</td>
<td>0.50</td>
<td>0.55</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>RoBERTa flow 0</td>
<td>0.44</td>
<td>0.54</td>
<td>0.52</td>
<td>0.55</td>
<td>0.35</td>
<td>0.43</td>
<td>0.42</td>
<td>0.47</td>
</tr>
<tr>
<td>RoBERTa flow 5</td>
<td>0.32</td>
<td>0.42</td>
<td>0.45</td>
<td>0.46</td>
<td>0.26</td>
<td>0.33</td>
<td>0.36</td>
<td>0.40</td>
</tr>
<tr>
<td>T5 flow 0</td>
<td>0.44</td>
<td>0.51</td>
<td>0.51</td>
<td>0.52</td>
<td>0.37</td>
<td>0.41</td>
<td>0.45</td>
<td>0.46</td>
</tr>
<tr>
<td>T5 flow 5</td>
<td>0.43</td>
<td>0.50</td>
<td>0.49</td>
<td>0.49</td>
<td>0.35</td>
<td>0.40</td>
<td>0.44</td>
<td>0.43</td>
</tr>
<tr>
<td>T5 flow 11</td>
<td>0.44</td>
<td>0.51</td>
<td>0.51</td>
<td>0.53</td>
<td>0.37</td>
<td>0.42</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>BERT mean</td>
<td>0.04</td>
<td>0.14</td>
<td>0.10</td>
<td>0.11</td>
<td>-0.03</td>
<td>0.11</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td>fine-BERT mean</td>
<td>0.03</td>
<td>0.09</td>
<td>0.05</td>
<td>0.03</td>
<td>-0.03</td>
<td>0.10</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>BERT-large mean</td>
<td>-0.01</td>
<td>0.20</td>
<td>0.10</td>
<td>0.28</td>
<td>-0.03</td>
<td>0.14</td>
<td>-0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>fine-BERT-large mean</td>
<td>-0.02</td>
<td>0.11</td>
<td>0.04</td>
<td>0.17</td>
<td>-0.09</td>
<td>-0.05</td>
<td>-0.12</td>
<td>-0.06</td>
</tr>
<tr>
<td>T5 mean</td>
<td>-0.00</td>
<td>0.06</td>
<td>-0.00</td>
<td>0.07</td>
<td>-0.02</td>
<td>0.10</td>
<td>0.02</td>
<td>0.19</td>
</tr>
<tr>
<td>E-Z Reader</td>
<td>0.64</td>
<td>0.65</td>
<td>0.69</td>
<td>0.67</td>
<td>0.46</td>
<td>0.51</td>
<td>0.56</td>
<td>0.56</td>
</tr>
</tbody>
</table>

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)