Relation Classification with Cognitive Attention Supervision

Erik S. McGuire
DePaul University
Chicago, IL
emcguir8@depaul.edu

Noriko Tomuro
DePaul University
Chicago, IL
tomuro@cs.depaul.edu

Abstract

Many current language models such as BERT utilize attention mechanisms to transform sequence representations. We ask whether we can influence BERT's attention with human reading patterns by using eye-tracking and brain imaging data. We fine-tune BERT for relation extraction with auxiliary attention supervision in which BERT's attention weights are supervised by cognitive data. Through a variety of metrics we find that this attention supervision can be used to increase similarity between model attention distributions over sequences and the cognitive data without significantly affecting classification performance while making unique errors from the baseline. In particular, models with cognitive attention supervision more often correctly classified samples misclassified by the baseline.

1 Introduction

For humans, the task of determining semantic relationships may entail complicated inference based on concepts' contexts (Yee and Thompson-Schill, 2016; Zhang et al., 2020) and commonsense knowledge (e.g., causal relations; Chiang et al., 2021), and for labeling relations between entities in texts the task may depend on the genre of the text (e.g. biomedical, biographical) and constraints indicated by annotator instructions (Mohammad, 2016). The advent of crowdsourcing for machine learning approaches to Natural Language Processing (NLP) creates challenges in collecting high quality annotations (Ramírez et al., 2020). A platform such as Amazon Mechanical Turk (MTurk) allows accessible, sophisticated task design (Stewart et al., 2017) but defaults to simple templates for NLP tasks, and is susceptible to self-selection bias (raters may not represent the population) and social desirability bias or demand effects, where judges seek to confirm the inferred hypotheses of experimenters (Antin and Shaw, 2012; Mummolo and Peterson, 2019; Aguinis et al., 2020).

Cognitive research has shown that self-reports are frequently inaccurate (Vraga et al., 2016), and that subjects are unable to effectively introspect about or recall their eye movements during reading (Võ et al., 2016; Clarke et al., 2017; Kok et al., 2017). This encourages the use of precise, objective recordings of non-conscious language processing behavior to use as model training data, rather than relying solely on reader annotations. As emphasized by Hollenstein et al. (2019), when reading humans produce reliable patterns that can be recorded, such as tracking gaze trajectories or measuring brain activity. These signals can associate linguistic features with cognitive processing and subsequently be applied to NLP tasks. The recording of eye movements during reading can be traced to psychology and physiology in the late 1800s (Wade, 2010), but the use of eye-tracking data in NLP is a relatively new phenomenon (Mishra and Bhattacharyya, 2018). Brain data has a longstanding relationship with language processing and in recent years has been investigated with NLP models (Schwartz et al., 2019), leveraged notably by Mitchell et al. (2008) to predict fMRI activity from novel nouns.

The working intuition in using cognitive data in recent NLP studies is that signals produced by humans during naturalistic reading can be leveraged by artificial neural networks to induce human-like biases and potentially improve natural language task performance. For example, recognizing and relating entities while reading sentences might elicit patterns of activation or particular gaze behaviors in human readers which can be transferred to and recovered by models given the same text sequences as inputs. Models might then generalize learned biases to similar text inputs. One route for augmenting neural networks with cognitive data is to regularize attention, such as with eye-tracking (Barrett et al., 2018) and/or electroencephalography (EEG) data (Muttenthaler et al., 2020). Eye-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 222–232
Online Event, June 10, 2021.

Phrase	Relation
<e> ford </e> became an engineer with the <e> edison illuminating company </e>	Employer
<e> ford </e> became an <e> engineer </e>	Job Title
<e> ford </e> was born on a prosperous farm in <e> springwells township </e>	Birthplace
<e> mary litogot </e> (c1839-1876) , immigrants from <e> county cork </e>	Visited

Table 1: Some example phrases for sentences 3 and 5.

tracking (ET) is an indirect estimate of processes such as attentional focus and cognitive strategies (Eckstein et al., 2017) by associating eye movements with performance; EEG is a direct measurement of brain activity, by recording the electric potentials along the scalp generated by the firing of populations of neurons. We focus in this work on deep learning-based approaches to NLP and seek to induce human-like biases in the self-attention distributions produced by BERT[1] (Devlin et al., 2019) by fine-tuning the base language model for relation classification (RC) with a multi-task learning (MTL) approach, supervising attention with ET and EEG data taken from the Zurich Cognitive Language Processing Corpus (ZuCo[2]; Hollenstein et al. 2018) as the auxiliary task.

2 Related Work

Mathias et al. (2020) describe the key terms used in gaze behavior studies; eye-tracking appears to be the more robust and proven measurement modality for augmenting machine learning models. In particular, **fixations** are the eyes' focused pauses on Areas of Interest (AOIs); **saccades** are rapid movements from one point to another. These movements can be progressive or regressive, moving to later or earlier AOIs (e.g., the words in a sentence), and occur on the order of milliseconds. Hollenstein et al. (2019) combine the indirect signals of ET with EEG data, moving beyond inferences based on eye-screen positioning (e.g., that content words are more likely to be fixated upon, and unfamiliar words have longer fixation durations). In general, EEG provides a high temporal resolution but due to interference from the scalp exhibits a poorer spatial resolution than other brain imaging methods such as magnetoencephalography (MEG; Hollenstein et al., 2020). To understand cognitive processes involved in, e.g., longer fixation durations, EEG can complement ET, where larger amplitudes for event-related potentials (ERPs) such as N400 correspond to less frequent or less predictable words and semantic processing (Frank et al., 2015).

A number of studies have applied cognitive data to NLP tasks, among them: sentiment analysis (Mishra et al., 2016), part-of-speech (POS) tagging (Barrett et al., 2016), and named entity recognition (NER) (Hollenstein and Zhang, 2019). Hollenstein et al. (2019) apply both gaze and brain data to a suite of NLP tasks (Hollenstein et al., 2019), including relation classification. For sentiment analysis, Mishra et al. (2018) use MTL for a bidirectional Long Short-Term Memory (biLSTM) network, learning gaze behavior as the auxiliary task. Malmaud et al. (2020) predict ET data with a variant of BERT as an auxiliary to question answering. Bautista and Naval (2020) predict gaze features with an LSTM to evaluate on sentiment classification and NER tasks. Barrett et al. (2018) supervise model attentions with ET data by adding attention loss to the main classification loss so the model jointly learns a sentence classification task and the auxiliary task of attending more to tokens on which humans typically focus. Muttenthaler et al. (2020) follow this paradigm using EEG data.

A number of studies impose schemata or mechanisms to encourage BERT to learn more structured RC representations: Soares et al. (2019) fine-tune BERT for RC, experimenting with the use of additional special entity tokens from BERT's final hidden states to represent relations, rather than the last layer's classification token, [CLS]: the [CLS] token is conventionally used as the sentence representation for tasks such as classification (Devlin et al., 2019), as well as attention analysis (Clark et al., 2019). For joint entity and relation extraction Xue et al. (2019) fine-tune BERT using focused attention to mask what the [CLS] token attends to, so that it attends only to entities. Su and Vijay-Shanker (2020) fine-tune BERT for RC by summarizing the other tokens' final hidden states with either LSTM or attention, concatenating the result to the [CLS] representation.

[1] https://huggingface.co/bert-base-uncased

[2] https://osf.io/2urht/

Relation	Train	Train %	Test	Test %	Total
Awarded	9	1.77%	1	1.75%	10
Birthplace	68	13.36%	8	14.04%	76
Deathplace	17	3.34%	2	3.51%	19
Education	36	7.07%	4	7.02%	40
Employer	31	6.09%	3	5.26%	34
Founder	13	2.55%	1	1.75%	14
Job Title	136	26.72%	15	26.32%	151
Nationality	38	7.47%	4	7.02%	42
Political Affiliation	13	2.55%	2	3.51%	15
Visited	129	25.34%	15	26.32%	144
Wife	19	3.73%	2	3.51%	21
Totals	509	100%	57	100%	566

Table 2: Statistics for the static, stratified train and test splits on 566 phrase samples derived from 300 ZuCo sentences, as a given sentence may contain multiple binary relations among entities.

3 Data

Hollenstein et al. (2018) created ZuCo, a corpus of ET and EEG recordings in which 12 adult subjects (fluent English speakers) read full sentences at their own speed, with brain recordings synchronized to eye fixations. The sentences used by the corpus were written English: 400 review excerpts from Stanford Sentiment Treebank (Socher et al., 2013) and 707 biographical sentences from a Wikipedia relation extraction dataset (Culotta et al., 2006). In this work we use a subset of 300 relation sentences (7,737 tokens) divided into 566 phrases[3] by Hollenstein et al. (2019) to encompass the multiple binary relation statements, and annotated with markers around entity mentions. The dataset uses 11 relation types, as seen in Table 2.

For ET we had access to five features for each word, including first fixation duration (FFD), gaze duration (sum of fixations), and total reading time (TRT: the sum of the word's fixations including regressions to it). The features for EEG we use are the 105 electrode values mapped to first-pass fixation onsets to create fixation-related potentials (FRPs), so that each word has 105 values. We average ET and EEG values over all subjects, which has been shown to reduce variability of results (Hollenstein et al., 2020) and overfitting (Bingel et al., 2016). To obtain a single ET value for each token, Barrett et al. (2018) used the mean fixation duration (MFD), by dividing TRT by number of fixations. There is no best practice to our knowledge, and in this study we use TRT as a proxy for overall attention to a word. For EEG electrode values, we obtain a scalar for each word by taking the mean (Hollenstein et al., 2019), rather than the maximum (Muttenthaler et al., 2020).

[3] https://github.com/DS3Lab/zuco-nlp/tree/master/relation-classification/data

4 Method

We split the English-language ZuCo samples into 90% training and 10% test sets. We perform 9-fold cross-validation on the training data for 6 epochs with batch size 16 and otherwise default hyperparameters, averaging validation results over folds. We fine-tune the final models on the full training data, choosing 4 epochs based on cross-validation accuracy, reserving the test data for later comparison. For the main RC task, categorical cross-entropy loss $\mathcal{L}_{RC}$ is calculated for each sequence j in batches of size M with sequence-level predictions for the C classes, $\hat{y} \in \mathbb{R}^{M \times C}$, and a vector of target class indices $t \in \mathbb{Z}^M$ where $0 \leq t_j < C$:

$$\mathcal{L}_{RC}(\hat{y}, t) = -\frac{1}{M} \sum_{j}^{M} \ln\ a_{t_j} \qquad (1)$$

where a_{t_j} is the t_j-th value of the softmax of sample j's C prediction scores $\varphi(\hat{y}_j)$:

$$a = \varphi(\hat{y}_j) \qquad a_{t_j} = \frac{e^{\hat{y}_{jt_j}}}{\sum_{k}^{C} e^{\hat{y}_{jk}}}$$

We additionally compute auxiliary attention losses. BERT takes an input of sequence hidden states $\in \mathbb{R}^{N \times d}$ (N tokens, $d = 768$ features) and uses 12 attention heads at each layer to create 12 token-token attention weight matrices $\in \mathbb{R}^{N \times N}$.

Model	Loss	Accuracy	Precision	Recall	Weighted F1
Baseline	0.61	**0.88**	**0.83**	**0.80**	**0.88**
ET	**0.60**	0.87	0.82	**0.80**	0.87
EEG	0.62	0.86	0.80	*0.77*	0.87
ET+EEG	0.63	0.86	0.82	**0.80**	*0.85*
Random ET	*0.64*	*0.85*	*0.78*	0.78	0.86
Random EEG	0.62	0.86	0.82	0.79	0.86
Random ET+EEG	0.62	0.87	**0.83**	**0.80**	0.86

Table 3: Metrics at 4 epochs, averaged over 4 runs. Bold are best values, italics worst. Weighted macro-F1 is intended to account for class imbalances.

Specifically, in these matrices, there is a row for every token in the sequence–a distribution of N attention weights, where each scalar weight corresponds to a token's similarity to a token in the sequence. The resulting matrices are multiplied with the input to transform the tokens' features and produce a context matrix $\in \mathbb{R}^{N \times d}$. Each token context vector c contains a blend of features from the sequence's tokens: each feature for c is a weighted sum dominated by that feature's values from tokens most attended by c. For instance, the features in the context vector for [CLS] will reflect the features of those tokens given highest attention by [CLS], with the features of lower weighted tokens scaled down and contributing minimally.

These operations are founded on the conception of attention emerging from relationships between tokens in sequence contexts, or the notion of each token attending to the others, and computations occur in the subspaces of heads' attention weights: this is incompatible with the concept of a single abstracted human reading a displayed word sequence. Therefore, to intervene on the production of contextualized model representations using the ZuCo data as proxies for attention, we seek a single distribution of weights from the multiple token-token attention matrices for a given sequence, analogous to the competitive attention given by a human reader. Due to its use as the sequence representation used for classification, we take from each matrix the row of weights accorded by [CLS], resulting in 12 vectors, treating [CLS] as our model reader. We average these vectors along the head axis to obtain a [CLS]-token vector $\alpha \in \mathbb{R}^{1 \times N}$ of attention weights. This aggregate is supervised during training: in this way, each independent representation subspace (head) is informed by the human values, influencing the features of the sequence representation used for the RC task.

We then obtain human scores for the sequence tokens. Previous studies used "type-aggregated" (Barrett et al., 2016; Hollenstein et al., 2019) cognitive data, where values are averaged over corpus word occurrences to obtain an aggregated value for that word type. This method exchanges specific sample contexts for the ability to synthesize distributions for samples not in the original data through type lexicon queries, using 0 for unknown word types. For relation extraction, previously Hollenstein et al. (2019) discretized and binned ZuCo features which were used in an auxiliary task. To preserve context, we extract from ZuCo the raw ET and EEG values for each sample without type-aggregating, so that ZuCo coverage of tokens in the samples is complete: every token has a ZuCo value, excluding special model tokens, which are assigned zeros.

Because BERT uses subword tokenization, to allow matching entries to be found in the ZuCo word-level data we split the ZuCo words into BERT tokens, evenly dividing values between each subword piece (e.g., "delicacy" $\rightarrow$ "del", "##ica", "##cy", each piece allotted a third of the ZuCo value), a technique used by Malmaud et al. (2020). We preserve entity markers "<e>" and "</e>" in each sample by adding them as special tokens to the BERT tokenizer so their embeddings are learned with other tokens during fine-tuning. Human ET and EEG token values z_{ET} and z_{EEG} are passed through softmax to obtain two distributions over sequences, vectors α''_{ET} and α'_{EEG}. ET features such as TRT are much larger, measured in milliseconds, than the small EEG microvoltages (μV), so the raw ET values' softmax output α'_{ET} would be much peakier than α'_{EEG}, providing an extremely low entropy signal where weights are forced onto one or two tokens. To combat this, we reduce each ET token value by dividing by the maximum value

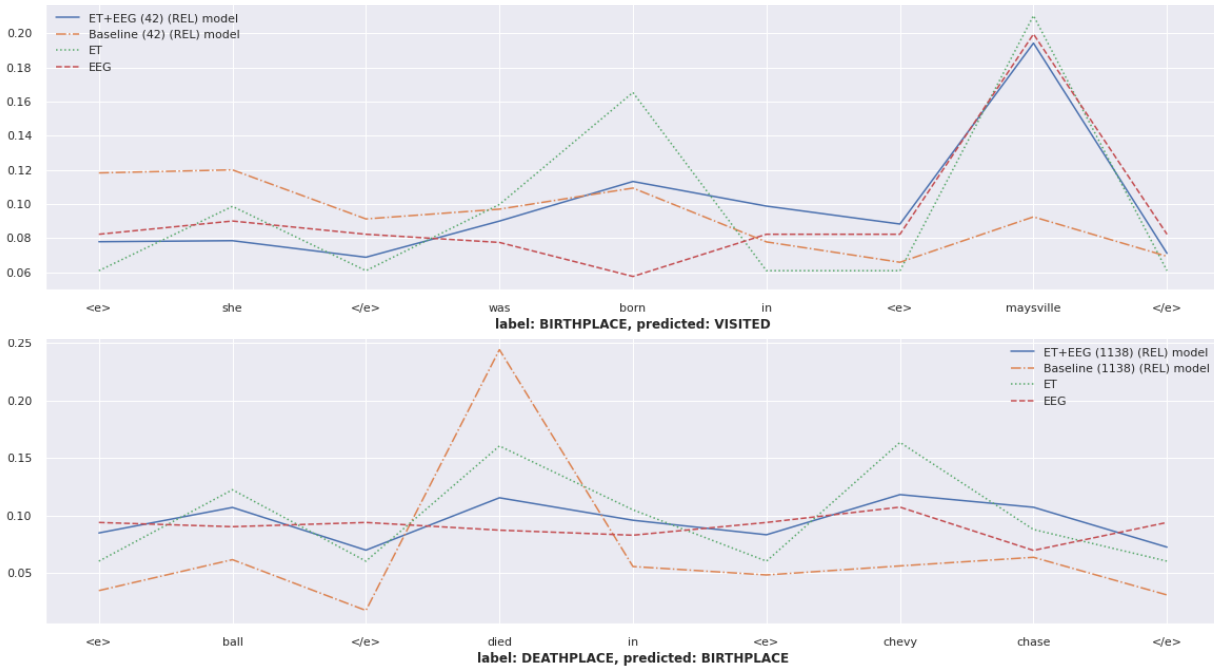

Figure 1: Plots of baseline and attention-supervised model attentions against ZuCo ET and EEG values where the baseline is correct and attention-supervised model is incorrect. Note that piece attentions are combined (e.g., "may": 0.1004 + "##sville" 0.0939 → "maysville": 0.1943). The ET+EEG model in the top plot was influenced to emphasize the location "maysville" alongside "born in" and predicts "Visited" rather than the correct "Birthplace", whereas the baseline places relatively more emphasis on "she was" and "born". At bottom, the baseline attends strongly to "died" whereas the ET+EEG model has learned a more uniform attention distribution.

for its sequence (Eq. 3), returning softmax output α''_{ET}. Each sequence thereby has a context-specific distribution, reflecting the averaged responses of the human subjects. Following other studies that implemented attention supervision (Qiuxia et al., 2020; Sharan et al., 2019; Sood et al., 2020; Zhang et al., 2019), we compute attention losses based on the Kullback-Leibler divergence (D_{KL}[4]) from the aggregate model attention weights α to the human weights α''_{ET} and α'_{EEG}. We do so for each sequence j in batches of size M for each modality, obtaining eye-tracking loss $\mathcal{L}_{ET}$ and EEG loss $\mathcal{L}_{EEG}$. By toggling binary coefficients λ, one or both losses are added to RC categorical cross-entropy loss to give us the overall multi-task fine-tuning loss, $\mathcal{L}_{MTL}$.

$$\mathcal{L}_{ET} = \frac{1}{M}\sum_{j}^{M} D_{KL}(\alpha_j''^{ET} || \, \alpha_j) \quad (2a)$$

$$\mathcal{L}_{EEG} = \frac{1}{M}\sum_{j}^{M} D_{KL}(\alpha_j'^{EEG} || \, \alpha_j) \quad (2b)$$

$$\mathcal{L}_{MTL} = \mathcal{L}_{RC} + \lambda_{ET}\mathcal{L}_{ET} + \lambda_{EEG}\mathcal{L}_{EEG} \quad (2c)$$

where $\alpha_j''^{ET}$ is the softmax of the max-normalized vector of ET token values for sequence j:

$$\frac{z_j^{ET}}{\max(z_j^{ET})} \quad (3)$$

5 Experimental Results

5.1 Ablations

We perform ablations comparing base BERT fine-tuned for four runs with arbitrary random seeds and varying combinations of the cognitive data. The baseline used in ablations is the result of fine-tuning on the ZuCo data without attention supervision. For the ET model, we add only the loss computed from the ET data. For the EEG model we do likewise with the EEG loss, and for the combined ET+EEG model we compute and add both auxiliary losses to the main classification loss. We similarly create random ET, EEG, and ET+EEG

[4]For this computation, zeros are set to 1e-12.

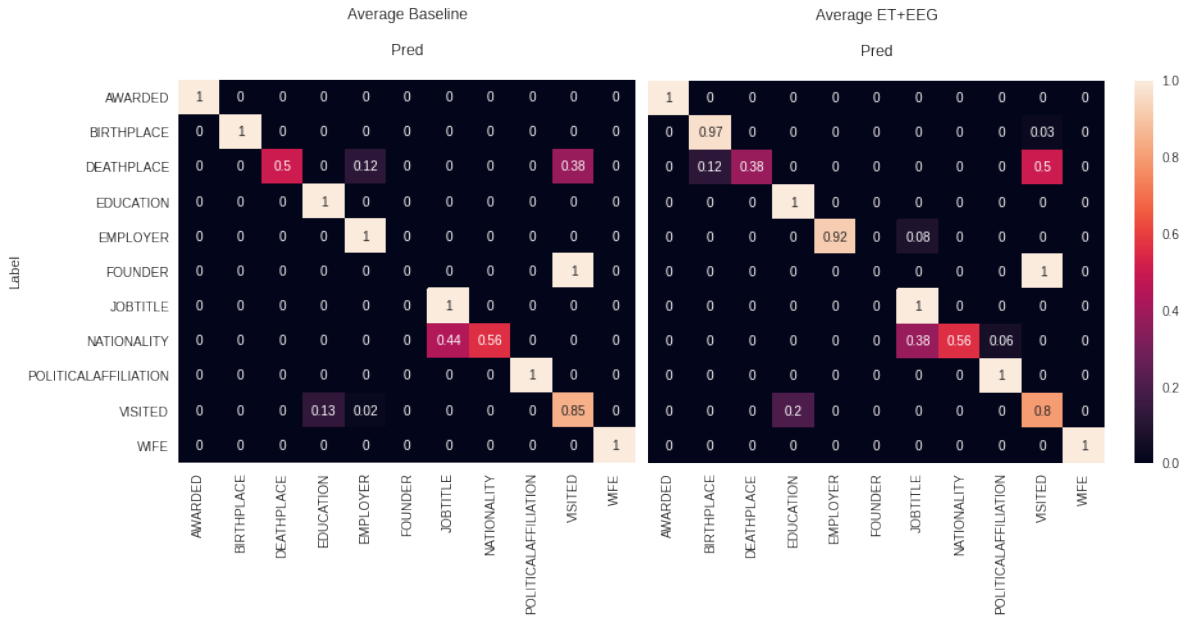

Figure 2: Confusion matrices of accuracies averaged over four runs for baseline and ET+EEG models. Models often misclassified "Deathplace" (which comprises roughly 3.5% of the splits' samples) as "Visited" (25%) or "Nationality" (7%) as "Job title" (26%), and "Visited" was occasionally misclassified as "Education" (7%).

models. For random models, we replace the modality's ZuCo values with values uniformly sampled from the fixed minimum and maximum range of the modality's ZuCo values. This should allow us to distinguish the effects of learning regularities in ZuCo token attention values versus the effects of constraining the range of magnitudes given by the ZuCo values.

After training, we evaluated the final models on the held-out test data of 57 samples. Table 3 shows evaluation results. Two-sided Pitman's permutation tests (Dror et al., 2018) were performed on final accuracies to assess statistical significance, comparing each of the six models against the baseline. Averaging over four runs, there are no statistically significant differences ($p > 0.05$) between baseline vs. ET, EEG, ET+EEG, and random versions thereof, respectively. Figure 2 displays confusion matrices for the models, showing similar per-class results, with some cases where classes with few samples such as "Deathplace" (19 samples) were classified as more dominant categories such as "Visited" (144 samples).

5.2 Attention Similarity

Sen et al. (2020) define a *behavioral similarity* metric to quantify the extent to which model attentions focus on the same words as the human attention; in their work, human attention maps are binary vectors used as the ground truth against which the continuous model attention maps are compared using Area Under the Curve (AUC), a binary classification metric. In a similar vein, in order to assess whether models learn a generalizable bias in attention we create a measurement to assess the amount of token overlap between continuous human and model attention vectors for phrases in the test set. Results of this measurement as well as relative entropies are shown in Table 4.

We compare a fixed top-k tokens for sequences using a variety of k values, for tokens scored by model attentions after fine-tuning and the scores given by human data. We run the models on all splits, using the methods described in §4 to obtain model attentions α, and compute the attention similarity for the test set by pairwise comparison of each model's attentions with the human data. Specifically, as Equation 4 describes, for each model we obtain sets of all samples' token indices and values for the top k attention weights from both ZuCo values α' and α, and divide the cardinality of the sets' intersection by k to obtain an overlap ratio. To factor the k weights' salience into the similarity, we divide their total weight given by the model by their total ZuCo weight and multiply this percentage–capped at 1.0–with the overlap ratio. For example, if both the baseline and an attention-supervised model have the same tokens in the top k

	k=1		k=2		k=3		k=4		D_{KL}	
Model	EEG	ET	EEG	ET	EEG	ET	EEG	ET	EEG	ET
Baseline	*4.26*	3.50	*8.23*	*7.70*	*12.90*	*13.69*	*20.10*	*17.37*	*0.36*	*0.44*
ET	12.11	**26.09**	**27.53**	**36.28**	32.57	**45.59**	36.35	**48.58**	0.05	**0.04**
EEG	9.48	*3.29*	16.38	9.09	21.58	16.06	25.49	20.33	**0.02**	0.07
ET+EEG	**14.21**	18.48	25.64	26.62	**33.21**	35.57	**36.61**	41.39	0.03	0.05
Random ET	11.92	6.82	20.61	16.99	29.80	29.28	33.88	33.79	0.04	0.05
Random EEG	13.11	9.29	17.50	18.27	27.77	30.38	34.68	36.69	0.06	0.07
Random ET+EEG	12.98	8.09	20.11	16.88	29.35	30.26	34.67	36.87	0.05	0.06

Table 4: Overlapping top-k and batch Kullback-Leibler divergence for model vs. sample-specific human attentions on the test set. Averaged over 4 runs, bold cells are best, italics worst.

attention, the model that weighs these tokens similarly to the ZuCo data should have a greater score. We take the average over each sample j in dataset D:

$$\text{sim}(\alpha_j, \alpha'_j) = \frac{1}{D}\sum_{j}^{D}\left[\frac{|o_j^k|}{k} \times \min\left(1, \frac{\sum_{i}^{o_j^k}\alpha_{ji}}{\sum_{i}^{o_j^k}\alpha'_{ji}}\right)\right] \quad (4)$$

where o_j^k is the set of intersecting indices of the top k attention values for sequence j and α' corresponds separately to α'^{ET} (Eq. 3) or α'^{EEG}:

$$o_j^k = \alpha_j^k \cap \alpha_j'^k \quad (5)$$

As Table 4 shows, baseline and random models have less overlap than the ET model for all sets. Curiously, after the baseline, EEG overlap was weakest for the model supervised with EEG, including for the random models. This might indicate a diffusion of attention that makes top-k overlap difficult to differentiate, as EEG overlap values reach parity with non-EEG models with $k > 10$. Figure 1 visualizes the respective final [CLS] attention weights averaged over attention heads for baseline vs. attention-supervised models against the ET and EEG ZuCO data values used to supervise the latter models.

5.3 Unique Errors

While task performance is not significantly different, we can see that model attentions are affected. To detect the possible effects of these attentional differences, where alternative features may be emphasized or diminished in the sequence representations used for RC, we analyze errors made by the baseline models against those of the attention-supervised models on a sample by sample basis. For each model $\mathcal{M}_a$ paired with baseline model $\mathcal{M}_b$ (fine-tuned without attention supervision), we examine the proportion of the pair's mismatched errors out of all errors on the test set (Equation 6); that is, the size of the symmetric difference ($\triangle$) between $\mathcal{M}_a$'s errors $\mathcal{M}_a^{\text{inc}}$ and $\mathcal{M}_b$'s errors $\mathcal{M}_b^{\text{inc}}$ divided by the size of the union of errors made by each model:

$$\text{mismatches}(\mathcal{M}_a, \mathcal{M}_b) = \frac{|\mathcal{M}_a^{\text{inc}} \triangle \mathcal{M}_b^{\text{inc}}|}{|\mathcal{M}_a^{\text{inc}} \cup \mathcal{M}_b^{\text{inc}}|} \quad (6)$$

As seen in Table 5, we note that models with non-random ZuCo attention supervision have more unique errors compared with the baseline than those with random supervision. In this case, the EEG-based attention loss seems to be the source of the small differences, as ET and Random ET models have similar mismatches. Lin et al. (2020)

Model	MM	Fixes	Breaks	AvB
Baseline	0.00	0.00	0.00	0.00
ET	20.59	0.07	*0.02*	*0.16*
EEG	24.16	**0.11**	*0.02*	*0.16*
ET+EEG	**28.90**	**0.11**	**0.04**	**0.22**
Random ET	*20.24*	*0.03*	0.03	0.18
Random EEG	20.83	*0.03*	0.03	0.18
Random ET+EEG	25.55	**0.11**	0.03	0.18

Table 5: *MM* (*mismatches*): The percentage of unique errors between model errors and baseline errors out of all errors for both models. *Fixes* refers to the percentage of all $\mathcal{M}_b$'s errors that $\mathcal{M}_a$ correctly predicted. *Breaks* refers to the percentage of all $\mathcal{M}_b$'s correct answers that $\mathcal{M}_a$ incorrectly predicted. *AvB* refers to the percentage of all $\mathcal{M}_a$'s errors that $\mathcal{M}_b$ correctly predicted. Bold cells are the highest, italicized lowest.

examine *fixes*: instances where the baseline is in error, but the modified baseline is correct. We analyze the percentages of *fixes* and also *breaks*, which we define to occur when the baseline is correct, but the model with supervised attention is incorrect. These are also shown in Table 5. Compared to random models, the ZuCo models seem to more frequently predict correctly samples that the baseline labeled incorrectly.

6 Conclusions and Future Work

Overall, BERT models with multiple modes of human attention supervision converged to accuracy for the relation classification task that does not differ significantly from the fine-tuned base BERT model, despite possessing attention distributions that were shifted toward the cognitive data. Measured by overlap, attention supervision with eye-tracking data was most influential on the final layer's [CLS]-assigned attention weights. In addition, we have shown that the behavior of these models differs from the baseline consistently by misclassifying different samples, exposing pathologies which may be of interest for research in neural network-based human language processing.

Barrett and Hollenstein (2020) have pointed to distinct reading patterns evident in eye-tracking studies for unfamiliar proper nouns which may be more readily apparent in the ET values. On the other hand, it may be that the EEG data were too noisy and that dimensionality reduction to find the most predictive electrode values, such as performed by Muttenthaler et al. (2020), is needed to provide a consistent signal. Additionally, Hollenstein et al. (2019) and Muttenthaler et al. (2020) incorporated EEG frequency bands into their ZuCo-based studies; the α frequency band has been associated with attention (Feldmann-Wüstefeld and Awh, 2020) and supervision with this band might yield different results. The cognitive data used in this study were not specifically produced from an entity-related reading task, but Brédart (2017) has noted the increased difficulty of processing proper names which is reflected in behavioral studies, with a double dissociation between common nouns and proper names where production of one type of noun is impaired but the other is intact. A more careful use of neuroimaging data may be needed to leverage signals reflecting the differing brain mechanisms involved in human lexical access.

Typically, researchers implicitly seek to induce a human-like bias in classifiers so they correlate more highly with human judgments by using self-reported annotations to supervise learning. This supervision is limited insofar as self-reports can not specify responses inaccessible to annotator introspection, such as the brain's electrical activity or detailed gaze behavior. Models additionally biased by non-conscious physiological responses may learn to more robustly reflect human language processing, incorporating both subjective and objective signals. Human annotations are conventionally taken as ground truth. Yet cognitive data may offer valid judgments, as well. For example, in sentiment analysis, a false negative according to a self-report could be a true negative according to physiological affective responses. Cognitive data may reveal inconsistencies and gradations obscured by labels. In the case of relation extraction, cognitive data might uncover patterns more reflective of different, potentially novel categories of semantic relation, or different dynamics, due to linguistic ambiguity and/or changing contexts and readerships. In terms of limitations, we did not investigate the breadth or depth of influence of our method of [CLS]-based aggregate attention supervision on the model attentions across layers and heads, nor the supervision of specific layers or heads as done by Strubell et al. (2018). We did not explore trade-off coefficients on the multiple losses, such as the convex combination used by Malmaud et al. (2020). We used a relatively small English dataset, which limited generalizability and robustness.

Hollenstein et al. (2020) describe some ethical concerns in the recording and use of cognitive data, including voluntary data procured but not recorded by NLP researchers. This includes loss of privacy with the identification of subjects, an overrepresentation and normalization of particular demographics, and the perpetuation of fossilized human prejudices. Sen et al. (2020) have described the potential for human attention supervision to address the validity of attention as a faithful, human-like explanation for model decisions while Pruthi et al. (2019) have discussed the potential for deception by manipulating attention to make models appear less biased. Future work could scrutinize whether human attention supervision can provide a basis for exploring cognitive biases learned by models, or align attention-based explanations to model outcomes: enabling performant models to adhere faithfully to auditor expectations.

References

Herman Aguinis, Isabel Villamor, and Ravi S Ramani. 2020. MTurk research: Review and recommendations.

Judd Antin and Aaron Shaw. 2012. Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2925–2934.

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.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–584, Berlin, Germany. Association for Computational Linguistics.

Maria Barrett and Nora Hollenstein. 2020. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for Natural Language Processing. *Language and Linguistics Compass*, 14(11):1–16.

Louise Gillian Bautista and Prospero Naval. 2020. Towards learning to read like humans. In *International Conference on Computational Collective Intelligence*, pages 779–791. Springer.

Joachim Bingel, Maria Barrett, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fMRI-with an application to PoS induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755.

Serge Brédart. 2017. The cognitive psychology and neuroscience of naming people. *Neuroscience & Biobehavioral Reviews*, 83:145–154.

Jeffrey N Chiang, Yujia Peng, Hongjing Lu, Keith J Holyoak, and Martin M Monti. 2021. Distributed code for semantic relations predicts neural similarity during analogical reasoning. *Journal of Cognitive Neuroscience*, 33(3):377–389.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alasdair DF Clarke, Aoife Mahon, Alex Irvine, and Amelia R Hunt. 2017. People are unable to recognize or report on their own eye movements. *The Quarterly Journal of Experimental Psychology*, 70(11):2251–2270.

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, New York City, USA. 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.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Maria K Eckstein, Belén Guerra-Carrillo, Alison T Miller Singley, and Silvia A Bunge. 2017. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental cognitive neuroscience*, 25:69–91.

Tobias Feldmann-Wüstefeld and Edward Awh. 2020. Alpha-band activity tracks the zoom lens of attention. *Journal of cognitive neuroscience*, 32(2):272–282.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing NLP with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

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.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-*

guage Technologies, Volume 1 (Long and Short Papers), pages 1–10, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellen M Kok, Avi M Aizenman, Melissa L-H Võ, and Jeremy M Wolfe. 2017. Even if i showed you where you looked, remembering where you just looked is hard. *Journal of Vision*, 17(12):2–2.

Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A Miller. 2020. Does BERT need domain adaptation for clinical negation detection? *Journal of the American Medical Informatics Association*, 27(4):584–591.

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.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharya. 2020. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4907–4913. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Abhijit Mishra and Pushpak Bhattacharyya. 2018. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking*. Springer.

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.

Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey. 2018. Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179, San Diego, California. Association for Computational Linguistics.

Jonathan Mummolo and Erik Peterson. 2019. Demand effects in survey experiments: An empirical assessment. *American Political Science Review*, 113(2):517–529.

Lukas Muttenthaler, Nora Hollenstein, and Maria Barrett. 2020. Human brain activity for machine attention. *arXiv preprint arXiv:2006.05113*.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*.

LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*.

Jorge Ramírez, Marcos Baez, Fabio Casati, Luca Cernuzzi, and Boualem Benatallah. 2020. Challenges and strategies for running controlled crowdsourcing experiments. *arXiv preprint arXiv:2011.02804*.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608, Online. Association for Computational Linguistics.

Komal Sharan, Ashwinkumar Ganesan, and Tim Oates. 2019. Improving visual reasoning with attention alignment. In *International Symposium on Visual Computing*, pages 219–230. Springer.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

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.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. 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.

Neil Stewart, Jesse Chandler, and Gabriele Paolacci. 2017. Crowdsourcing samples in cognitive science. *Trends in cognitive sciences*, 21(10):736–748.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. pages 5027–5038.

Peng Su and K. Vijay-Shanker. 2020. Investigation of BERT model on biomedical relation extraction based on revised fine-tuning mechanism.

Melissa L-H Võ, Avigael M Aizenman, and Jeremy M Wolfe. 2016. You think you know where you looked? you better look again. *Journal of Experimental Psychology: Human Perception and Performance*, 42(10):1477.

Emily Vraga, Leticia Bode, and Sonya Troller-Renfree. 2016. Beyond self-reports: Using eye tracking to measure topic and style differences in attention to social media content. *Communication Methods and Measures*, 10(2-3):149–164.

Nicholas J Wade. 2010. Pioneers of eye movement research. *i-Perception*, 1(2):33–68.

K. Xue, Y. Zhou, Z. Ma, T. Ruan, H. Zhang, and P. He. 2019. Fine-tuning BERT for joint entity and relation extraction in chinese medical text. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 892–897.

Eiling Yee and Sharon L Thompson-Schill. 2016. Putting concepts into context. *Psychonomic Bulletin & Review*, 23(4):1015–1027.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. 2020. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications*, 11(1):1–13.

Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. 2019. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 349–357. IEEE.