Eye Gaze and Self-attention: How Humans and Transformers Attend Words in Sentences

Anonymous ACL submission

Abstract

Attention mechanisms are used to describe human reading processes and natural language processing by transformer neural networks. On the surface, attention appears to be very different under these two contexts. However, this paper presents evidence that there are links between the two during reading tasks. During reading, the dwell times of human eye movements were strongly correlated with the attention patterns occurring in the early layers of pre-trained transformers such as BERT. Furthermore, we explored what factors lead to variations in these correlations and observed that data were more correlated when humans read for comprehension than when they were searching for specific information. Additionally, the strength of a correlation was not related to number of parameters within a transformer.

1 Introduction

Attention is highly associated with reading in humans and with Natural Language Processing (NLP) by state-of-the-art Deep Neural Networks (DNN) (Bahdanau et al., 2014). In both cases, it is the words within a sentence that are attended during processing. For humans, this attention process is strongly linked to eye gaze (Rayner, 2009), with words at the center of an eye fixation being the words that are attended. In DNNs, this attention process is the result of mechanisms built into the network. In the case of the current state-of-the-art method Transformers (Vaswani et al., 2017), this attention process is the result of the dot product of two vectors of that represent individual words in the text.

While attention in human reading processes and transformers appear to be completely different, this paper will present experimental evidence of a link showing the relationship between the two. Specifically, the attention in well-known transformers such as BERT (Devlin et al., 2019), and its derivatives are closely related to humans' eye movements during reading. We observed strong to moderate strength correlations between the dwell times of eyes over words and the self-attention in transformers such as BERT. We have explored some reasons for these different correlation levels but note a general strong link between the original BERT model and the movements of the human eye.

1.1 Transformers

Since their introduction, Transformers (Vaswani et al., 2017) have dominated the leader boards for NLP tasks. They have also impacted computer vision (Dosovitskiy et al., 2020), including generative networks (Jiang et al., 2021). The transformers primary feature is the attention mechanism,

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{n}}\right)\mathbf{V} \tag{1}$$

Word vectors representations from the sequence Q are compared to those from sequence K. This is used to determine how much information word representations from the former should incorporate from the latter. If the query and key sequence are the same, as in a transformers encoder, it is called self-attention. The results of the attention process are then multiplied by sequence V to get the final outputs from the attention layer. V contains different representations for the words in K.

The more important K words are to those in Q, the more attention Q words allocate to that word. The Q x K part of the attention mechanism has been examined to understand how transformers process information. Vaswani et al. (2017) showed that words in Q could learn anaphora resolution by appropriately attending the word "its" in the K.

The proliferation of pre-trained models quickly followed the introduction of transformers. Arguably, the most famous of these models is BERT, a.k.a. the Bidirectional Encoder Representations from Transformers model (Devlin et al., 2019). BERT is used to encode information from whole passages of text into a single vector. Its bidirectional structure means that each token is placed in the context of the entire sequence instead of just the words appearing before it. This structure provided an increase in performance on the GLUE benchmarks (Wang et al.) over mono-directional models such as the original GPT (Radford et al., 2018).

To ensure that the model learned to attend to the sequence as the whole, BERT was trained using Masked Language Modeling (MLM), a task inspired by Cloze procedure (Taylor, 1953) from human reading comprehension studies. In MLM, random words from a sequence were hidden during input. The model then has to predict what word was hidden based on the context of the surrounding words. Additionally, BERT was trained to perform Next Sentence Prediction (NSP), forcing words from one sentence to attend to words in other sentences. BERT achieved state-of-the-art performance in multiple NLP benchmarks following this training regime, which led to its fame.

BERT's impact on the field can be seen in the number of subsequent models that are its direct descendants. Examples include models such as RoBERTa (Liu et al., 2019), which uses BERT's architecture but was trained via different methods. Other models, such as ALBERT (Lan et al., 2019), were created to condense BERT for faster performance with minimal accuracy loss. Even models such as XLNet (Yang et al., 2019) extend BERT's architecture to include recurrence mechanisms introduced in other models (Dai et al., 2019). In turn, some of these descendant models have been used to create other models. For example, BIGBIRD (Zaheer et al., 2020) was built using RoBERTa as its base model.

1.2 Combining Transformers and Eye Gaze

There is a growing field of research that combines pre-trained transformers with eye-tracking data. Researchers have used BERT outputs as features for machine learning models to predict eye fixations. In some instances, these outputs are combined with other features (Choudhary et al., 2021), whereas in other instances, BERT itself is fine-tuned to perform the task. For example Hollenstein et al. (2021a) have shown that BERT can be effective at predicting eye movements for texts written in multiple languages, including English, Dutch, German, and Russian.

Given the strong relationship between eye gaze and attention, it is unsurprising that there have been recent attempts to compare eye gaze to the attention generated in transformers. Sood et al. (2020a) compared eye movements in reading comprehension task to three different neural networks, including XLNet. After fine-tuning XLNet, they compared the attention from the last encoder layer to eye gaze and reported a non-significant correlation. With that said, their comparison only reported the correlation for the final attention layer of the network, and other studies comparing transformer attention to human metrics have indicated that the strength of an association can differ by layer (Toneva and Wehbe, 2019). Therefore, the present study look at all of the layers of the transformers.

Following the work of Sood et al. (2020a), the present study is a large-scale analysis of the relationship between attention in pre-trained transformers and human attention derived from eye gaze. We compared the self-attention values of 31 variants from 11 different transformers, including BERT, its descendants, and a few other state-of-the-art models (Table 1). Using BERT-based models allowed us to investigate what effect the training regime has on how closely the attention is related to eye-based attention. Using non-BERT models allowed us to examine what effect model architecture has on this relationship. Finally, the different datasets enabled an investigation into how the humans' task also affects this relationship. Results showed a surprisingly strong correlation between attention in the first layer of the transformers and total dwell time. These correlations were unrelated to the size of the model.

2 Related Work

There have been attempts to combine DNNs with eye data to perform various tasks. Some basic tasks include predicting how an eye will move across presented stimuli, whether text-based (Sood et al., 2020b) or images in general (Ghariba et al., 2020; Li and Yu, 2016; Harel et al., 2007; Huang et al., 2015; Tavakoli et al., 2017). These predictions can be used to create saliency maps that show what areas of a visual display are attractive to the eye.

In turn, the saliency maps can be used to either understand biological visual processes or be incorporated as meta-data into machine learning models. The later process has led to some improvements

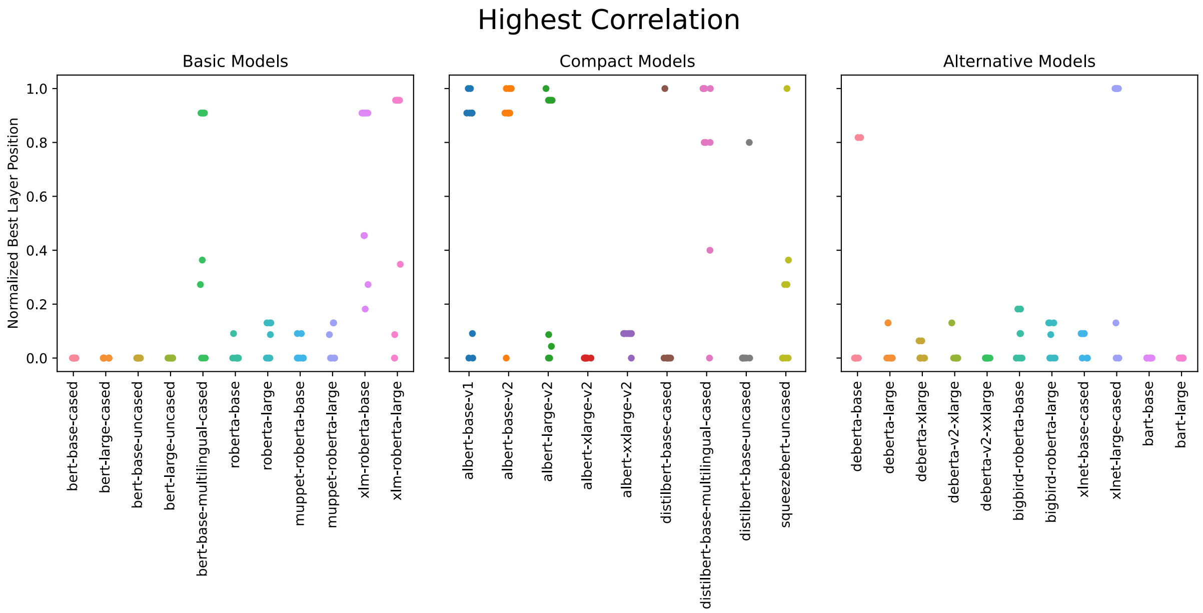


Figure 1: The relative position of the layer with the highest correlation. 0 is the first layer, 1 is the last layer. Each dot represents the highest correlation from one dataset.

in task performance. In a recent example, Sood et al. (2020b) achieved state-of-the-art results in a text compression task by creating Text Saliency Model (TSM) using a BiLSTM network that output embeddings into transformer self-attention layers. The TSM was pre-trained on synthetic data simulated by the E-Z reader model (Reichle et al., 1998) and fine-tuned on human eye-tracking data. The model's output was used to neuromodulate (Vecoven et al., 2020) a task-specific model via multiplicative attention.

Alternatively, eye data itself can be used to inspire new ways for neural networks to perform NLP tasks (Zheng et al., 2019). For example, it is well known that the human eye does not fixate on every word during reading (Duggan and Payne, 2011). Nevertheless, humans, until recently, performed well above machines in many NLP tasks (Fitzsimmons et al., 2014; He et al., 2020). These observations imply that the word skipping process is not detrimental to reading-based tasks. Some researchers have exploited this process by explicitly training their models to ignore words (Yu et al., 2017; Seo et al., 2018; Hahn and Keller, 2016). For example, Yu et al. (2017) trained LSTM models to predict the number of words to skip while performing sentiment analysis and found that the model could skip several words at a time and still be as accurate, if not more accurate, than the non-skipping models. Additionally, Hahn and Keller (2018) showed that you could model the skipping processes using actual eye movements and achieve the same result.

Another type of exploration between DNNs and human data is to examine how closely the metrics used to measure eye movement are related to metrics used for machine language models. Studies of this type require identifying comparable processes between the two different systems and a suitable dataset. For example, Ettinger (2020) has tested model performance with human concepts such as commonsense knowledge or negation, while Hao et al. (2020) has compared model perplexity to psycholinguistic features.

There have even been comparisons of DNN attention to what humans attend to during reading tasks. Sen et al. (2020) compared the attention of humans during a sentiment analysis task to RNN models. Crowdsourced workers were asked to rate sentiments of YELP reviews and then highlight the important words for their decision-making process. They found correlations between the RNN outputs and human behavior. The strength of these correlations diminished as the length of the text increased.

Closely related to this is Sood et al. (2020a) who attempted to compare eye gaze to the attention mechanisms of three different neural network architectures. One of the models was the BERT-based transformer, XLNet (Yang et al., 2019). The

Table 1: List of models used in this paper

Model	Pretrained models in Huggingface repository
ALBERT (Lan et al., 2019)	albert-base-v1, albert-base-v2, albert-large-v2, albert-xlarge-v2, albert-xxlarge-v2
BART (Lewis et al., 2019)	facebook-bart-base, facebook-bart-large
BERT (Devlin et al., 2019)	bert-base-uncased, bert-large-uncased, bert-base-cased, bert-large-cased, bert-base-multilingual-cased
BIGBIRD (Zaheer et al., 2020)	google-bigbird-roberta-base, google-bigbird-roberta-large
DeBERTa (He et al., 2020)	microsoft-deberta-base, microsoft-deberta-large, microsoft-deberta-xlarge, microsoft-deberta-v2-xlarge, microsoft-deberta-v2-xxlarge
DistilBERT (Sanh et al., 2019)	distilbert-base-uncased, distilbert-base-cased, distilbert-base-multilingual-cased
Muppet (Aghajanyan et al., 2021)	facebook-muppet-roberta-base, facebook-muppet-roberta-large
RoBERTa (Liu et al., 2019)	roberta-base, roberta-large
SqueezeBERT (Iandola et al., 2020)	squeezebert-squeezebert-uncased
XLM (Conneau et al., 2020)	xlm-roberta-base, xlm-roberta-large
XLNet (Yang et al., 2019)	xlnet-base-cased, xlnet-large-cased

other two networks were bespoke CNN and LSTM models. All models were trained on the MovieQA dataset (Tapaswi et al., 2016), and attention values were taken from the later levels of the networks. Several questions for the original dataset were selected for human testing, where the participants' eye gazes were tracked while they read and answered the questions. Sood et al. (2020a) observed that the attention scores from both the CNN and LSTM networks had strong negative correlations with the eye data. However, there was no significant correlation between eye gaze and XLNet.

Finally, there has been recent work using transformer representations to predict brain activity. For example, Toneva and Wehbe (2019) used layer representations of different transformers, including BERT and Transformer-XL, to predict activation in areas of the brain. They found that the middle layers best predict the activation as the context (sequence length) grew. Toneva and Wehbe (2019) tentatively suggested that this means there is a relationship between the layer and the type of processing occurring. To their surprise, they also found that changing lower levels of BERT to produce uniform attention improved prediction performance.

Schrimpf et al. (2020) performed a similar analysis using many of the models included in the present study. They found that the output of some transformers could be used to predict their participants brain behavior to almost perfect accuracy. Prediction performance differed by model size and training regime, with GPT-2 performing best (Radford et al., 2019). Surprisingly, Schrimpf et al. (2020) found that untrained models also produced above chance prediction, leading them to suggest that the architecture of transformers captures import features of language before training occurs.

3 Analysis of Self-Attention Against Eye Gaze

All analyses used HuggingFace's (Wolf et al., 2020) version of the transformer and associated tokenizer. No training was conducted on any of the models. Data were created by inputting tokenized sequences into the transformer and extracting the attention matrices produced for each attention head. In terms of Equation 1, we are taking the output of the softmax function before it is multiplied by V as that provides a normalised value indicating what proportion of attention each token pays to all others. The attention value for each token was calculated by averaging across attention heads and matrix rows. This calculation produced a single vector representing the amount of attention allocated to each token by all others in the sequence. Finally, if a word was tokenized into sub-words, those sub-words were also averaged to produce a single value for the entire word. The special tokens [CLS] and [SEP] were used for the attention calculations but dropped from the final word level attention vector.

Our procedure differs from Sood et al. (2020a) who used the maximum attention from each word instead of the mean. Our experiments showed that the mean attention values provided more stable results across datasets. For comparison purposes, the results using the maximum values have been provided in the supplemental data.

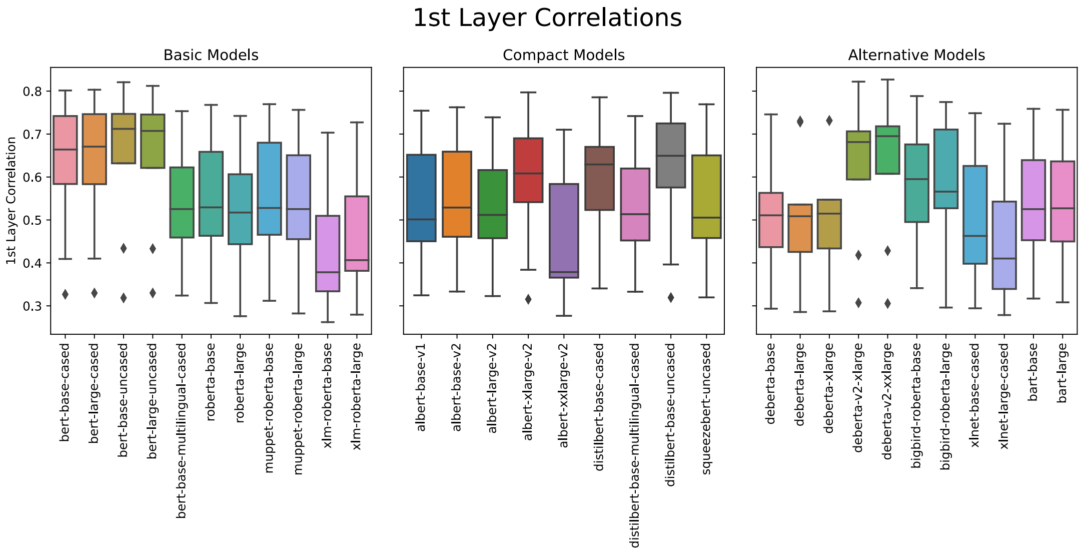


Figure 2: The correlations between the first layer attention patterns and eye-tracking data from all datasets.

3.1 Datasets and Models

Six different datasets were used in this analysis. In all cases, eye-tracking data were captured from human participants performing reading based tasks in English. Full descriptions these dataset can be found in Appendix A.

Our experiments used 31 variants from 11 different transformers models. We have grouped these variants into three types: 1) Basic models use the same architecture as BERT. 2) Compact models are those designed to be smaller versions of basic models. 3) Alternative models are those that greatly differ from the basic models. A full list of models and variants can be seen in Table 1, further details can be found in Appendix B.

3.2 Results and Discussion

All analyses reported here refer to the average dwell time for each word. Dwell time is a measurement of the total time that a participant's eye fixated on a word. Both the eye gaze data and transformer attention outputs were normalized by sentence. Final values refer to the proportion of either dwell time or attention output during the processed sentence. All analyses report Spearman correlations (Coefficient, 2008) to avoid data normality issues and provide a direct comparison to previously reported work.

There were significant positive correlations between the total dwell time and the attention from all layers of the different models. This finding was an apparent departure from the results of Sood et al. (2020a) who reported a non-significant correlation of -.16 between the last layer of XLNet and their dataset. For comparison, we obtained a .428 correlation for their Study 1 data and .327 for their Study 2 data from XLNet's last layer. Although they did not directly specify the normalization they used, we suspect that the difference in results is due to us using sentence level normalization and Sood et al. (2020a) using paragraph normalization. For comparison, we ran the same procedure using paragraph normalization and obtained non-significant correlations just as they did. Many of the correlations obtained in sentence-level comparisons become much weaker at the paragraph level. This finding corresponds well with Sen et al. (2020) finding that attention for non-transformer neural networks becomes less correlated with eye movements as the length of the text increases. Due to this, all analyses presented here refer to sentence-level correlations.

Our first analysis investigated which attention layer was most closely correlated with the human data. Figure 1 shows the relative position of the layer with the highest correlation by model. In many cases, the highest correlation was produced by the earlier layers of each model, usually the first layer (position 0). Notable exceptions to this rule are the multilingual versions of BERT and RoBERTa (i.e., XLM) and many compact models. The finding that multilingual variants of models do not behave like monolingual variants is in line with previously reported studies (Conneau et al., 2020; Hollenstein et al., 2021b; Vulić et al., 2020), where some studies report benefits and others not.

Further investigations found that when the first

Table 2: First layer correlations By dataset. Strongest correlations have been bolded.

Model	GECO	Mishra	Provo	Sood S1	Sood S2	ZuCo S1	ZuCo S2	ZuCo S3	Frank et al
albert-v1	0.744	0.754	0.497	0.450	0.326	0.501	0.580	0.325	0.652
albert-v2	0.748	0.739	0.492	0.460	0.329	0.503	0.585	0.326	0.637
bart	0.729	0.758	0.526	0.451	0.323	0.511	0.550	0.313	0.638
bert-cased	0.802	0.783	0.668	0.584	0.410	0.643	0.679	0.328	0.744
bert-multilingual-cased	0.753	0.727	0.525	0.459	0.338	0.489	0.622	0.324	0.603
bert-uncased	0.816	**0.791**	**0.710**	**0.626**	**0.434**	**0.693**	**0.722**	0.324	**0.746**
birdbird-roberta	0.775	0.774	0.600	0.511	0.363	0.582	0.565	0.319	0.693
deberta-v1	0.731	0.735	0.511	0.432	0.310	0.502	0.533	0.289	0.549
deberta-v2	**0.824**	0.770	0.708	0.601	0.423	0.688	0.712	0.306	0.660
distilbert-cased	0.786	0.772	0.623	0.523	0.378	0.629	0.632	0.341	0.670
distilbert-multilingual-cased	0.742	0.740	0.513	0.452	0.337	0.487	0.620	0.333	0.602
distilbert-uncased	0.796	0.780	0.649	0.576	0.396	0.649	0.678	0.319	0.725
roberta	0.709	0.755	0.523	0.453	0.329	0.504	0.537	0.291	0.632
roberta-muppet	0.712	0.763	0.527	0.460	0.329	0.501	0.542	0.297	0.665
squeezebert	0.730	0.769	0.505	0.458	0.320	0.499	0.549	**0.348**	0.650
xlm	0.690	0.715	0.391	0.358	0.271	0.379	0.476	0.313	0.532
xlnet	0.678	0.736	0.436	0.369	0.287	0.408	0.470	0.297	0.584

layer did not produce the highest correlation, the first layer value was close to the best value to represent the performance of the model. An extreme example of this were the ALBERT variants, which, likely due to weight sharing during training, have virtually identical correlations from attention values from each of its levels (Figure 3). Due to its general best performance, the first layer results have been used at the best performance for all models. Analyses using the actual best performance can be observed in the supplemental files, although those results are highly similar to those reported here.

Our next analysis compared performance across models based on the 1st layer correlations. Figure 2 shows that, in general, the size of the model does not determine the correlation between the human eye and transformer attention. Evidence for this can be seen in minor differences between different sized variants of the same model. For example, the cased and uncased versions of BERT-base and BERT-large are very similar, despite the large variants' containing 340 million parameters compared to the base variants' 110 million. Similar observations can be observed across the other models, especially DeBERTa, where the largest variants have 1.5 billion parameters, and the smaller ones contain less than 1/3 of that number. Due to this similarity, results in Table 2 reports a single value per model type that is an average for each size variant. Table 3 shows the highest correlation by dataset. In most cases, this model was either BERT-uncased or DeBERTa-V2.

While the number of parameters is not what determines the correlations, comparing across models in Figure 2 suggests that training is essential for determining those relationships. For example, the BERT models have identical architectures to various RoBERTa models, yet Table 2 shows that the BERT correlations were consistently higher than the RoBERTa based models. The other clear examples of training effects can be seen in the differences between DeBERTa V1 and V2, where V2 models use the Scale-invariant-Fine-Tuning (SiFT) algorithm introduced in the original paper. Interestingly, the addition of the SiFT algorithm allowed DeBERTa V2 to surpass human performance on the SuperGLUE benchmarks (Wang et al., 2019), and Table 3 shows that this model was often the second-highest correlated model. While it would be great to find a direct relationship between how humanlike a model's performance is and how correlated its attention patterns are to eye movements, that is not the case. Excluding the compact models, the BERT descents outperform it on many of the benchmarks, yet only DeBERTA comes close to having stronger correlations to human eye movements. In most cases, attention patterns less related to human attention produce better overall performance on NLP tasks.

Tables 3 and 2 show the rankings by correlation are similar between datasets, with BERT-uncased producing the highest correlation in all but two cases. In one of the exceptions, the GECO dataset, BERT-uncased, was ranked second. In the other exception, ZuC0 Task 3, the ranking was much lower. In general, the correlations from ZuCo Task 3 differ greatly from the other datasets. The correlations

Table 3: The three models with strongest correlation to eye-tracking data for each dataset. The uncased version of BERT produced the strongest correlation in 7 out of 9 cases.

	GECO	Mishra	Provo	Sood S1	Sood S2	ZuCo S1	ZuCo S2	ZuCo S3	Frank-et-al
1	deberta-v2	bert-uncased	bert-uncased	bert-uncased	bert-uncased	bert-uncased	bert-uncased	squeezebert	bert-uncased
2	bert-uncased	bert-cased	deberta-v2	deberta-v2	deberta-v2	deberta-v2	deberta-v2	distilbert-cased	bert-cased
3	bert-cased	distilbert-uncased	bert-cased	bert-cased	bert-cased	distilbert-uncased	bert-cased	distilbert-multilingual	distilbert-uncased

are lower for all models, and the model rankings are very different, with two of the compart models, SqueezeBERT and DistillBERT, ranking highest, and BERT-uncased, ninth. Task 3's participants were the same as Tasks 1 and 2. Those first two tasks produced results closer to the other datasets, meaning Task 3's lower correlations are likely due to the task itself.

Interestingly, in Task 3, the participants were presented with the question on the screen, allowing them to direct their eye gaze to find the information they required. This allowance contrasts with most of the other datasets where the questions about the data were presented after reading. The only exceptions to this were some tasks by Sood et al. (2020a) where the question appeared on screen in Study 2 and in 2/3's of the tests in Study 1. Furthermore, the correlations from Sood et al. (2020a) Studies 2 and 1 were also the second and third lowest of the datasets, respectively (Table 2). While further study is needed, the lower correlations from SOOD et al. and ZuCo Task 3 may indicate that while transformer attention patterns produce strong correlations when reading typically, the relationship drops when the reader actively searches for information.

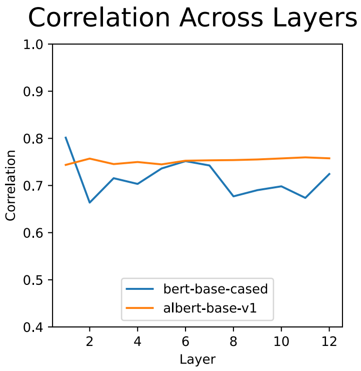


Figure 3: The average correlations between the attention patterns of across layers.

Our final analysis looks at correlations across levels of BERT (Figure 3). The results of Toneva and Wehbe (2019) suggest that the middle layers of BERT provided the best features for predicting brain activity in humans. They speculated that these relationships could mean that the middle layers of BERT could be related to the kinds of processing that occurs in those brain levels. Our results show that the attention patterns from BERTs first layer were closely related to eye gaze data. Again, while speculative, our results combined with Toneva and Wehbe (2019) would suggest that for BERT at least, the lower levels correspond best to text information entering the eyes. In contrast, the middle layers correspond to specific processing. With that said, not all transformers produced the strongest correlations from their first layer.

4 Investigating the Effect of Injecting Eye-Gazing Bias During Training

In this section, we investigated the effect of injecting human eye-gazing bias during training on test accuracy. We used the BERT model (Devlin et al., 2019) and the sarcasm-detection dataset published in Mishra et al. (2016) as a case study in our experiments.

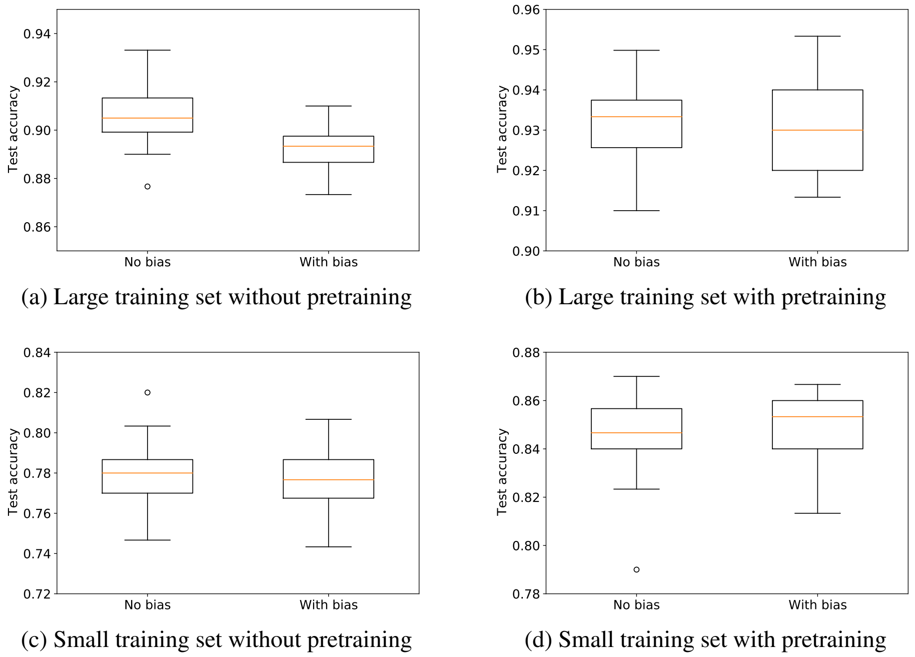
4.1 Method

The Mishra et al. (2016) dataset was originally proposed to predict non-native English speakers' understanding of sarcasm by using eye-tracking information. The dataset contains information on the fixation duration of each word for each participant. We injected the eye-gazing bias during training by optimising the following loss function

$$L = H(y, \hat{y}) + \alpha H(p, \hat{p}) \qquad (2)$$

where $H(y, \hat{y})$ is the cross-entropy loss of the binary classification task of sarcasm detection, and $H(p, \hat{p})$ computes the divergence of the first-layer attention values from the distribution of the normalised fixation duration values given a sentence. The hyperparameter α controls the weight of the second term in the loss function.

In our experiments, we only used the fixation duration values from participant 6 in the dataset because participant 6 had the highest overall accuracy



(a) Large training set without pretraining

(b) Large training set with pretraining

(c) Small training set without pretraining

(d) Small training set with pretraining

Figure 4: Comparison of the BERT models trained with eye-gazing bias against the models trained without in terms of test accuracy. Models in plots (a) and (b) were trained on 693 examples, and the results were obtained after 20 runs. Models in plots (c) and (d) were trained on only 70 examples, and the experiments were repeated 50 times. The same test set (300 examples) was used for all the experiments.

on sarcasm detection (90.29%). All the hyperparameters were tuned on a validation set extracted from the training set before they were applied on the full training set.

4.2 Results

The results of the experiments are plotted in Figure 4. As expected, the models fine-tuned from pretrained BERT models had significantly better test accuracy for both small training set and large training set than models trained from scratch on the Mishra et al. (2016) dataset.

When the models were trained on the large training set without pretraining, the injection of human eye-gazing bias during training actually hurt the performance (statistically significant using t-test under 0.05 confidence level). With pretraining, both models in Figure 4(b) performed better than the best participant in the Mishra et al. (2016) dataset. The injection of human bias still lowered the mean accuracy, although the difference was not statistically significant anymore. When the small training set was used to train the models, we found no significant difference after the injection of eye-gazing bias.

Comparing our results to Sood et al. (2020b) suggests that training a model to predict eye gaze improves text compression performance, whereas using eye gaze data to regulate sarcasm detection decreased performance. It is unknown whether the difference in results is due to our task choice or to our method of using human data.

5 Conclusion

This paper analyzed the correlations between attention in pre-trained transformers and human attention derived from eye gaze. We found correlations between the two that were generally stronger in the earlier layers of the model, and in most cases, strongest in the first layer. These correlations were unaffected by the model's size, as different sized variants of models produced similar correlations. The training the models received did appear to matter, although the present study cannot determine the full extent of that relationship. We found that correlations were weaker from eye-tracking studies where the participants could actively guide their reading towards seeking the information they needed than when they were presented with questions after reading. Finally, we showed that forcing the model's first-layer attention values to match the human attentions using eye-gazing duration data during training did not improve the model's performance on a sarcasm detection dataset.

Acknowledgements

References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *arXiv preprint arXiv:2101.11038*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Shivani Choudhary, Kushagri Tandon, Raksha Agarwal, and Niladri Chatterjee. 2021. Mtl782_iitd at cmcl 2021 shared task: Prediction of eye-tracking features using bert embeddings and linguistic features. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 114–119.

Spearman Rank Correlation Coefficient. 2008. The concise encyclopedia of statistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL (1)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Geoffrey B Duggan and Stephen J Payne. 2011. Skim reading by satisficing: evidence from eye tracking. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1141–1150.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Gemma Fitzsimmons, Mark Weal, and Denis Drieghe. 2014. Skim reading: an adaptive strategy for reading on the web.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45(4):1182–1190.

Bashir Muftah Ghariba, Mohamed S Shehata, and Peter McGuire. 2020. A novel fully convolutional network for visual saliency prediction. *PeerJ computer science*, 6:e280.

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.

Michael Hahn and Frank Keller. 2018. Modeling task effects in human reading with neural attention. *arXiv preprint arXiv:1808.00054*.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954*.

Jonathan Harel, Christof Koch, and Pietro Perona. 2007. Graph-based visual saliency.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021a. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021b. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123.

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.

Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 262–270.

Forrest Iandola, Albert Shaw, Ravi Krishna, and Kurt Keutzer. 2020. SqueezeBERT: What can computer vision teach NLP about efficient neural networks? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 124–135, Online. Association for Computational Linguistics.

Yifan Jiang, S Chang, and Z Wang. 2021. Transgan: Two pure transformers can make one strong gan, and that can scale up. CVPR.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Guanbin Li and Yizhou Yu. 2016. Visual saliency detection based on multiscale deep cnn features. *IEEE transactions on image processing*, 25(11):5012–5024.

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*.

Steven G Luke and Kiel Christianson. 2018. The provo corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Keith Rayner. 2009. The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology*, 62(8):1457–1506.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, NANCY G KANWISHER, Joshua B Tenenbaum, and Evelina Fedorenko. 2020. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv*.

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.

Minjoon Seo, Sewon Min, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Neural speed reading via skim-rnn. In *International Conference on Learning Representations*.

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 Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. *arXiv preprint arXiv:2010.07891*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Hamed R Tavakoli, Ali Borji, Jorma Laaksonen, and Esa Rahtu. 2017. Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features. *Neurocomputing*, 244:10–18.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32:14954–14964.

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.

Nicolas Vecoven, Damien Ernst, Antoine Wehenkel, and Guillaume Drion. 2020. Introducing neuromodulation in deep neural networks to learn adaptive behaviours. *PloS one*, 15(1):e0227922.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Adams Wei Yu, Hongrae Lee, and Quoc Le. 2017. Learning to skim text. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1880–1890.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human behavior inspired machine reading comprehension. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 425–434.

A Datasets

The GECO Corpus (Cop et al., 2017) contains data from 19 Dutch bilingual and 14 English readers who read "The Mysterious Affair at Styles" by Agatha Christie across four sessions, with comprehension tests occurring between sessions. The bilingual participants completed two sessions in English, two in Dutch. For our analysis, we selected all English sessions, regardless of the participant's bilingual status.

The PROVO Corpus (Luke and Christianson, 2018) contains 55 passages that contain an average of 2.5 sentences. Passages were taken from online news articles, magazines, and works of fiction. Eye-tracking data was captured from 84 native English speakers who were instructed to read for comprehension.

The ZuCo Corpus (Hollenstein et al., 2018) is a combined reading, eye-tracking, and EEG dataset. Data was captured from 12 native English speakers who could read at their own pace with sentences presented one at a time. The participants completed three different tasks. Task 1 was a sentiment analysis task. Task 2 was a standard reading comprehension task where participants were presented with questions after reading the text. Task 3 was also a reading comprehension task; however, the question appeared onscreen while the participant was reading.

We also used data from Sood et al. (2020a). They collected data from 32 passages taken from the MovieQA (Tapaswi et al., 2016) dataset. In Study 1, 18 participants answered questions from 16 passages under varying conditions such as multi-choice, free answer with text present, and free answer from memory. In Study 2, 4 participants answered multi-choice questions from the remaining 16 passages.

Additionally, we used data from Frank et al. (2013). In this set, 48 participants read 205 sentences from unpublished novels for comprehension. The dataset contains eye movements from a mix of native and non-native English speakers. Participants were occasionally required to answer yes/no questions following a sentence.

The final dataset comes from Mishra et al. (2016) who conducted a sarcasm detection task. The dataset was taken from a wide variety of sources, all short passages containing a maximum of 40 words. The eye gaze of non-native English speakers who were highly proficient in English was tracked while completing the task.

B Models

B.0.1 Basic Models

BERT (Devlin et al., 2019): On release, BERT was the state-of-the-art. It was trained using MLM, in which 15% of tokens were masked. Training also incorporated NSP by forcing the model to predict whether two sentences were contiguous or not. Our analysis includes both cased and uncased versions of the English BERT as well as a multilingual model.

RoBERTa (Liu et al., 2019): A Robustly Optimized BERT Pretraining Approach. The model's architecture is identical to BERT. However, RoBERTa was trained for longer, with larger batch sizes and more data. Unlike BERT, the MLM examples were dynamically generated during a batch, which used the same mask patterns every time a sample was used. Finally, the NSP task was dropped as it did not affect performance.

We have also included the MUPPET version of RoBERTa (Aghajanyan et al., 2021), trained using multitask learning with tasks from four domains: classification, commonsense reasoning, reading comprehension, and summarization. Finally, we have included XLM-RoBERTa (Conneau et al., 2020), a multilingual version of RoBERTa.

B.0.2 Compact Models

ALBERT (Lan et al., 2019): A Lite BERT is a BERT-based model that uses two tricks to reduce the number of parameters and time taken required to create the model. The first was factorized embedding parameterization. By decomposing the large vocabulary embedding matrix into two small matrices, the hidden size of embedding space can be different from the hidden size and much smaller. The second trick was cross-layer sharing. The parameters for all layers are shared, leading to faster training times.

DistilBERT (Sanh et al., 2019): This model used a Teacher – Student method for the distillation of knowledge (Buciluǎ et al., 2006; Hinton et al., 2015). Sanh et al. (2019) started with a full model and kept every second layer to create the student. The student was then trained on original training data. This procedure resulted in a model that was almost as powerful but half the size.

SqueezeBERT (Iandola et al., 2020): SqueezeBERT is Bert but with grouped convolutional layers instead of feed-forward layers. The model was trained using the same methods as ALBERT.

B.0.3 Alternative Attention Mechanisms

DeBERTa (He et al., 2020): Decoding-Enhanced BERT with Disentangled Attention differs from others on this list in that it decouples attention by word semantics from attention by word location. Version 2 of the model used a form of adversarial training to improve model generalization and surpassed human performance on Super GLUE benchmarks. We have used the RoBERTa base version of the model variants.

One problem with transformers is the quadratic memory, and computational growth as sequence length increases due to every token attending to all other tokens. Some have dealt with this problem by modifying the attention patterns to approximate this full attention pattern without requiring all of the attention comparisons. BIGBIRD (Zaheer et al., 2020) is an example that uses this attention approximation. The model uses a combination of global, sparse, and random attention. Again, we have used the RoBERTa version of the model.

B.0.4 Alternative Architectures

XLNet (Yang et al., 2019): This model is a BERT extension using random permutations of word order during training. The model also incorporates the recurrence mechanism used in Transformer-XL (Dai et al., 2019).

BART (Lewis et al., 2019): Bidirectional and Auto-Regressive Transformer is an encoder-decoder model that is to recover data from corrupted text input. BART has approximately 10% more parameters than comparable BERT models and no final feed-forward layer. Pre-training was based on corrupting the inputs using token masking, token deletion, token infilling, sentence permutation, and document rotation.