Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension

Anonymous submission

Abstract

While neural networks with attention mechanisms have achieved superior performance on many natural language processing tasks, it remains unclear to which extent learned attention resembles human visual attention. We study the similarity between human visual and neural attention and analyze if neural attention-based methods perform better if they mimic human attention. To this end, we compare state-of-the-art networks based on long short-term memory (LSTM), convolutional neural (CNN) and XLNet Transformer architectures on a question answering task. We evaluate all methods on a novel 23-participant dataset of eye tracking data recorded while reading movie plots. We find that while higher similarity to human attention and performance significantly correlates to the LSTM and CNN this does not hold true for the XLNet – despite the fact that the XLNet performs best on this challenging task. Our work not only shows that different architectures seem to learn rather different neural attention but also that similarity of neural to human attention is not necessarily helpful and hence desirable.

1 Introduction

Due to the high ambiguity of natural language, humans have to detect the most salient information in a given text and allocate a higher level of attention to specific regions to successfully process and comprehend it (Poesio, 1995; Shiffrin and Schneider, 1977; Schneider and Shiffrin, 1977). Eye tracking studies have been widely used in various reading comprehension settings to reveal these attentive strategies (Rayner, 2009) and have, as such, helped to interpret cognitive processes and behaviors during reading.

Recently, how the human attentive system works has inspired attention mechanisms in neural networks (Bahdanau et al., 2014; Hassabis et al., 2017). Like with humans, attention mechanisms allow networks to focus and allocate more weight to relevant parts of the input sequence (Vaswani et al., 2017; Xu et al., 2015; Chorowski et al., 2015; Mnih et al., 2014). As such, neural attention can be viewed as a model of visual saliency that makes predictions over the elements in the network's input, being a region in an image or a word in a sentence (Frintrop et al., 2010). Attention mechanisms have gained significant popularity recently, and have boosted performance in natural language processing tasks but also other fields such as computer vision (Seo et al., 2016; Veličković et al., 2017; Sun and Fisher, 2003; Ma and Zhang, 2003).

Contrary to the sophisticated NLP advancement, system performance degrades when models are exposed to some inherent properties of natural language, such as semantic ambiguity, inferring information, or out of domain data (Niven and Kao, 2019; Blohm et al., 2018). These findings encourage work towards enhancing network's generalizability, deterring reliance on the closed-world assumption (Reiter, 1981). It has been proposed, specifically within machine reading comprehension (MRC), the more similar systems are to human behavior, the more suitable they become for such a task (Trischler et al., 2016; Luo et al., 2019; Zheng et al., 2019). As a result, much recent work aims to build machines which read and understand with human-level performance (Blohm et al., 2018; Nguyen et al., 2016; Rajpurkar et al., 2016; Hermann et al., 2015). To that end, by employing self-attention, researchers attempt to enhance comprehension by building models which capture better deep contextual and salient information (Devlin et al., 2018; Zhang et al., 2018; Shen et al., 2018; Vaswani et al., 2017; Yu et al., 2018).

As neural attention allows us to "peek" inside neural networks, it can help us better understand

how humans make predictions. We leverage this to investigate the relationship between neural performance and attention similarities to humans. Addressing the human-like proposal, by interpreting and comparing the attention of three state of the art MRC models, our research questions are the following: (i) Is there any correlation between a particular network's performance and its similarity to human visual attention? (ii) Do attention models achieve state-of-the-art results on machine learning tasks in natural language processing because machine attention emulates human attention?

To answer these questions we extend a QA dataset - the MovieQA dataset (Tapaswi et al., 2016) - with eye tracking and present a novel visualization tool[1] to observe the real-time reading of humans versus models, in split screen mode. We extract human attention via gaze data which has been used to interpret the inner workings of the human mind (Lipton, 2016; Rouse and Morris, 1986; Van Hooft and Born, 2012; Milosavljevic and Cerf, 2008; Wiegreffe and Pinter, 2019) and interpret the relationship between three state-of-the-art systems for this dataset namely CNN, LSTM, and XLNet (Hochreiter and Schmidhuber, 1997; Yang et al., 2019) using Kullback-Leibler Divergence (Kullback and Leibler, 1951). By doing so, we are able to compare and better understand neural attention behaviors on text across attention models. To the best of our knowledge, we are the first to compare neural attention to human gaze data **on text based tasks**.

The main findings of our work are two-fold: First, we show that there is a statistically significant correlation within the CNNs and LSTMs model performances and similarity to human attention. Second, we show statistical significance that the LSTMs are more similar to humans attention, when compared to the XLNets, whereas they perform best on the MovieQA dataset.

2 Related Work

2.1 Eye-tracking for Attention and Comprehension

Eye tracking studies have been extensively used in cognitive science research to investigate human attention over time (Rayner, 1998). Importantly, it has been demonstrated that attention and saccadic movements must be intertwined (Deubel et al., 2000; Kristjansson, 2011; Hoffman and Subramaniam, 1995). Identification of attentional focus and eye movement can be evoked given intricate information processing tasks such as reading (Posner et al., 1980; Posner, 1980; Henderson, 1992); this is the connection between eye-tracking data, attention and reading.

Just and Carpenter (1980) developed a theory of reading comprehension (*The Reading Model*) which was used as a basis for the cognitive theory in this paper, in order to better understand the relationship between eye fixations, attention, and reading comprehension. In their eye tracking study, they measured cognitive processing load via fixation duration. They found that participants look longer or more often at items that are cognitively more complex, in order to succesfully process them. Cognitive load increases when readers are "accessing infrequent words, integrating information from important clauses and making inferences at the ends of sentences" (Just and Carpenter, 1980). This is the connection between attention and reading comprehension tasks.

2.2 Attention Mechanisms

In the attention-based encoder-decoder architecture, rather than ignoring the internal encoder states, the attention mechanism takes advantage of these weights to generate a context vector, which is used by the decoder at various time steps (Bahdanau et al., 2014; Luong et al., 2015; Chorowski et al., 2015; Wang and Jiang, 2016; Dzendzik et al., 2017; Yang et al., 2016).

In transformer networks, the main differences to previous attentive models, are that these networks are purely based on attention (there are not LSTM or GRU units), and attention is applied via self-attention and multi-headed attention (Vaswani et al., 2017). Since the introduction of pre-trained transformer networks, we have seen a rise in state-of-the-art performance across a multitude of tasks in NLP (Devlin et al., 2018; Yang et al., 2019; Radford et al., 2018). Given this, much effort has recently gone into interpreting these highly complex models (Vig, 2019).

2.3 Question Answering and Machine Comprehension

We use question answering tasks to evaluate human versus machine attention. Question answering tasks have been widely explored with neural

[1] This tool can be used to qualitatively interpret the differences and similarity in attentive behaviors between any neural and human black boxes.

attentive models. Creating systems to comprehend semantically diverse text documents and answer related questions is still a difficult challenge (Qiu et al., 2019). These models tend to fail when faced with adversarial attacks, noise which humans can often easily resolve (Jia and Liang, 2017; Yuan et al., 2019; Blohm et al., 2018). These studies uncovered the limitations of QA systems, suggesting that models rely on pattern matching in lieu of human decision making processes which are required in comprehension tasks (Blohm et al., 2018; Posner et al., 1980; Just and Carpenter, 1980). These models comprehend text differently than humans.

2.3.1 Eye Tracking and Neural Networks

In the past few years, researchers leveraging human gaze data for attentive neural modeling tasks. Hahn and Keller (2018) present a neural QA network which combines both a task and attention module to predict and simulate human reading strategies. The authors propose the *trade-off hypothesis:* human reading behaviors are task specific and therefore evoke various specific strategies for each of these tasks. To validate their hypothesis, they use eye tracking data as their gold standard. Das et al. (2017) investigate the differences between neural and humans attention **over image regions**, given a visual question answering task (i.e answering textual questions about a given image). The authors use rank-order correlation and visualizations in their analysis.

Recent work has even explored integrating gaze data into neural attention as an additional variable in the equation or as a regularization method (Qiao et al., 2018; Sugano and Bulling, 2016; Barrett et al., 2018).

2.4 Neural Interpretability

In order to further understand the black box, research in neural interpretability has grown dramatically in the recent years (Lipton, 2016; Gilpin et al., 2018). Such methods include: introducing adversarial examples, error class analysis, modeling techniques (e.g. self-explaining networks), and post-hoc analysis (Alvarez-Melis and Jaakkola, 2018; Rudin, 2019; Lipton, 2016).

Many works have shed light on the decisions taken by networks by investigating the outputs/predictions as well as by analyzing their behavior through loss visualization from various architectures (Ribeiro et al., 2016). However these interpretations might explain predictions without explaining the mechanisms by which models work (Lipton, 2016). There is still a limit to how we can interpret the inner workings of these black boxes (Gilpin et al., 2018).

3 Resources

3.1 MovieQA Dataset

The MovieQA dataset (Tapaswi et al., 2016), is used for all experiments conducted in this work (for the QA models and eye tracking experiments).

The dataset was comprised by a variety of available sources, however for the QA task we only use the plot synopses. The authors crawled for the plot synopsis on Wikipedia, retrieved the scripts from IMDB (which were used for about half the movies), and a small percentage of plot information comes from the DVS transcriptions. The plots vary between 1 to 20 paragraphs in size, and are checked by annotators to ensure they consist of movie relevant events and character relationships. There are a total of almost 15,000 questions in this dataset relevant for 408 movie plots. Of the 5 answer candidates denoted for each question, there is only 1 correct answer and the rest are deceptive incorrect answers. The training set consists of plots with their corresponding questions: 9,848 training, 1,958 development and 3,138 test questions, respectively.

3.2 Extension with Eye Tracking

We present a novel reading comprehension eye tracking dataset, for open use, which depicts how answering a question in various conditions evokes various comprehension strategies - indicated by eye movement differences in 3 conditions for the same document. Given the design of the eye tracking experiment, the dataset allows researchers to observe changes in reading behavior in three comprehension tasks, induce processing strategies evoked by humans, and provides a gold standard to compare and synchronize model versus human attention in comprehension tasks. We build and use our reading comprehension gaze dataset, as the gold standard, to further advance neural network interpretability in machine comprehension tasks.

Data collection Our corpus contains two studies: in Study 1 we randomly select a set of 16 documents in which the majority of both LSTMs and CNNs models failed to correctly answer the questions; in Study 2 we select a different set of 16

documents in which the majority of models succeed in predicting the correct answers.

In total, our corpus contains gaze data from 23 English native speakers who were recorded while reading 32 documents (around 200-250 words each) in three different comprehension tasks.

We used the Tobii 600Hz head-mount eye-tracker. In total, each session last 45 minutes including the time required for calibration and 5-minutes breaks every 15 minutes.

Study 1 16-documents were read by the participants in 3 different reading comprehension conditions. Our 3 conditions are designed as such: 1) regular QA where the subjects have access to the plot, the question, and 5 answer candidates; 2) open-ended answer generation where the subjects see the plot and the question but have to generate their own responses; and 3) QA by memory where the participants can first read the plot and then answer to the question (5 possible answers) without having the plot available. In condition 3, participants have to recover information from memory in order to answer the question. Participants are randomly distributed among the different conditions: 5-5-6 in schema A, 5-6-5 in schema B, and 6-5-5 in schema C. In order to maintain that every participant had the same number of data-points and that documents were seen in various conditions (avoiding a bias effect), we created 3 schemes and randomly assigned participants into these schemes. In schema A participants saw conditions ordered as C1, C2, C3, in schema B ordered as C2, C3, C1, and in schema C orders as C3, C1, C2. There are 16 documents, all were made for all three condition types (such that, for example, participant A in schema C read document 66 in condition 3, but participant D in condition 1 read the same document 66 in condition 1). This way we can see how the same task of generating an answer, in various conditions, evoked various comprehension strategies indicated by eye moment difference in conditions for the same document; hence our study is in accordance with a Latin Square Design.

Study 2 We conduct a follow up study in which we took only the plots that both the majority of CNN and LSTM models predicted correctly. We hypothesize that such items that are, on average, easier for the models are also easier for the humans (higher correlation score). In this study, we only collect data for the regular QA task (condition 1). The experiment was performed by 5 new participants. Each participant saw all the 16 plots in a randomised order.

[2]

Study	Schema	No. Doc	No. Participants	IAA	Acc
Study1	A	5	1-6	83.3%	93%
Study1	B	5	6-12	100%	100%
Study1	C	6	12-18	100%	100%
Study2	No-Schema	16	5	89.0%	95%

Table 1: Distribution in MovieQA with Eye Tracking.

Data agreement We only use data from the regular QA task as we have an equal number of data samples for both Study 1 and 2. (where we can compare attention and performance for difficult versus easy cases). Across both studies, we use 23 participants data total as they had the highest performance and agreement; accuracy and Pairwise Inter-annotator agreement was measured by Cohen's Kappa. The agreement is between 89-94% and the average accuracy is 95%.

4 Neural Models

4.1 Two Staged Attention Models

In this work, we re-implement both the CNN and LSTM QA ensemble models with two staged attention from Blohm et al. (2018) that provides state-of-the-art results on the MovieQA dataset (Tapaswi et al., 2016). These models are based on the compare-aggregate framework that achieves 85.12% on the test set and 84.37% on the validation set. In the multiple choice QA task, each datum contains the plot of a movie as well as its corresponding question and 5 potential answer candidates.

In the hierarchical structure, the models compare the plot to the respective question and aggregates this comparison into one vector representation to obtain a confidence score after applying the softmax, for each answer candidate. The best results were obtained from the majority vote of the nine best performing models.

The two-staged attention is performed twice, on the word and sentence level, where the plot is weighted with respect to the question or a possible

[2]To note, we have a reduced number of overall participants for study 1 however we do have the same amount of data samples as the participants read 16 documents for the same condition each.

answer candidate.

$$G = \text{softmax}\left(X^T P\right) \quad (1)$$

$$H = XG \quad (2)$$

The word level X indicates the answer candidate (5 total) or the question. Subsequently, when computing sentence level attention, the question or answer candidate are represented as such. Blohm et al. (2018) apply the dot-product computation for the attention mechanism. The two variations of this model with CNN and LSTM models provided state-of-the-art results on the MovieQA dataset with an average of 84.5% on validatation set and an average of 85.0% on evaluation set.

The authors perform a case study to further investigate the comprehension limitations of the models compared to human inference. In their analysis, they compared both networks against human performance in order to infer processing strategies which human possess but are shown by the models. They investigate the most difficult cases, where the majority of both 9-best models failed to correctly answer the question. This motivates why we use the difficult and easy documents for the CNN and LSTM models (Blohm et al., 2018), as they are the only paper to date which both as SOTA results and offered qualitative analysis on the gap between human and model performance. We deem difficult for the network via performance, when the majority of the models fail to correctly answer the question, we classify these documents as difficult cases for the two networks; vice versa for the easy documents.

4.2 XLNet Models

We use the pre-trained XLNet model, and fine-tune it for our QA task (Yang et al., 2019; Tapaswi et al., 2016). XLNet is a recently released transformer network for language understanding, which achieves state-of-the-art results on many NLP tasks (Yang et al., 2019). It was trained on large corpora with training objectives which are compatible with unsupervised learning and can be fine-tuned to new tasks and datasets.

XLNet is based on an auto-regressive approach in which the model uses observations from previous time steps in order to predict the weight for the next time step. Advancing from the traditional auto-regressive approach, such as a Bidirectional LSTM, the authors also combine their network with an auto encoding approach seen with the BERT model (Devlin et al., 2018). By combining both approaches, XLNet introduces permutations on both sides. Moreover, the self-attention network (Vaswani et al., 2017) uses three components, queries, keys and values, all of which are calculated from their respective embeddings. The output is a weighted sum of the values, in which the values are weighted with a score calculated as the dot product of the respective queries and keys. Its is important to note that the queries are related to the output and the keys are related to the given input.

During fine-tuning, however, the model is essentially the Transformer-XL (Dai et al., 2019; Yang et al., 2019; Vaswani et al., 2017). The auto regressive language model estimates the joint probability over the input elements (in XLNet this x is language agnostic, i.e it is a subtoken).

$$P(X) = \prod_t P(x_t | X_{<t}) \quad (3)$$

The input sequence is the concatenation of each x in the plot with the question and a potential answer candidate (there are 5 possible answer candidates and one correct answer).

We are fine-tuning on the task of question answering, where the model objective is multi-label classification given an input sequence. Note, the permutation language model is the component which helps XLNet capture longer dependencies between elements in a given input sequence (Yang et al., 2019). In our method, we fine-tune the XLNet with 24 attention layers and 16 attention heads (Yang et al., 2019). The fine-tuned model makes a prediction by applying the argmax over the softmax, selecting the potential y-label, or answer candidate, with the highest confidence scores.

Finetuning pretrained XLNet outperforms all other results on the validation set, obtaining the new highest accuracy of 91%[3].

5 Method

5.1 Evaluation Metric

In order to compare the human and neural attention distributions, we identified the Kullback-Leibler Divergence (Kullback and Leibler, 1951) to be the most suitable comparison method.

This method is used to compare two probability distributions, akin to relative entropy. We need to

[3] in progress for evaluating on the test-set to be added to the Leaderboard

ensure we compare the two distributions in a specific and consistent order; meaning we can compare either H to M or M to H, but we cannot interchange this direction. Additionally, the information we gain from this measure is an understanding of the differences in probability distribution between two variables (cf. Equation 4).

$$D_{\mathrm{KL}}(H \parallel M) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{H(x)}{M(x)} \right). \quad (4)$$

Concretely, we calculate the KL divergence for average-human to average-model along the word level attention distributions.

5.2 Human Gaze-Attention Extraction

We convert the raw gaze counts into a probability distribution by dividing each gaze count by the sum of all gaze counts. These token level frequency counts obtained in our hit testing method, reflect gaze duration; the more often a token of the text is attended to, the more important it is for humans to correctly answer the question (Just and Carpenter, 1980).

We extract word level attention weights and average over documents, thereby doing document level comparison given the word attention, because for humans the task is to look at the entire short document and then answer the question given the entire context, all items within the context are interconnected, and it would be misleading to only analyze attention over 1 sentence/parts of the document. To that extent, is not cognitively plausible to limit comparison to sentences or part of the documents as we given human access to the entire context. Therefore we compare attention given the entire context, i.e word attention weights over each document.

5.3 Extracting LSTM and CNN Word Level Attention

The sentence level attention for the CNN and LSTM models have very low entropy, where essentially almost all of the attention is distributed to one sentence and the rest of the sentence attention weights are almost 0. Given this is a property of the two-staged attention, which XLNet does not have. Therefore, we only analyze word level attention across humans and the three model types.

During evaluation, we extract token attention weights for each of the 9-best models. We then average the neural attention weights, given the selected answer candidate, across the same 32 plots used in the eye tracking study. We extract the attention from the selected answer candidate to ensure comparability to human attention. From the human data, we can only obtain attention given the answer they selected, thus to keep the neural vs. human attention comparable, we also only extract neural attention weights given the selected answer candidate.

5.4 Extracting XLNet Word and Sentence Level Attention

The attention weights from 9-best XLNet models are extracted from the output of the last hidden layer which contains token level weights for each plot-answer candidate pairing. In order to make the 1024-dimensional attention weights comparable to the human gaze attention, we only take the maximum value of each token attention (Htut et al., 2019) and normalize them by the sum of the weights we obtain this way.

6 Results

6.1 Analysis Results — Within Models

9-Best Model	Val Accuracy	Spearman	p−value
LSTM	84.37%	**-0.73**	< 0.001
CNN	82.58%	**-0.72**	< 0.001
XLNet	91.00%	-0.16	0.381

Table 2: Correlation within each model and performance, ensembles.

In Table 2, we show the within model type analysis. We report majority vote ensemble accuracy scores for each of the 9-best models, Spearman's rank correlation coefficient against human scores, and the corresponding p-values. Though ensembling methods are generally used to boost performance, for our experiments we performance analysis over ensembled networks for each type as well. We assert that just as we average each participants attention scores for a given document, we also average the models for each type, treating the 9-models for each model type as 9 participants (9 LSTM participants, 9 CNN participants, and 9 XLNet participants).

There are two **statistically significant negative correlation** results within the traditional attentive **LSTM and CNN models**, -0.73 and -0.72 with

$p < 0.001$, respectively. These correlation results indicate that for each of the 9-best model types, as the performance in answering each document-question correctly increases, the divergence to human visual attention decreases.

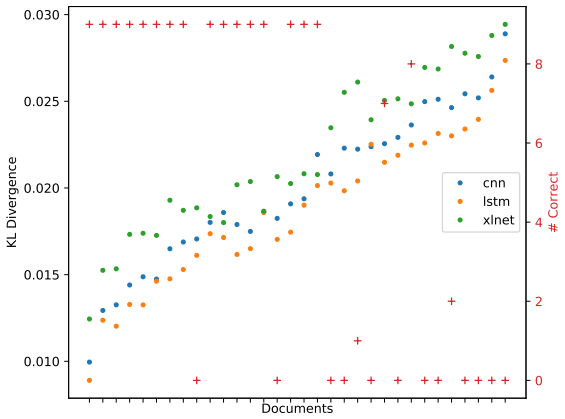

Figure 1: KL Divergence(Word Level): LSTM - All/Each Documents Ordered by the Sum of Divergence Scores and Number of Correct Models.

These correlations can be seen in Figures 1, where we plot the LSTM model performance for each document. The same behavior is observed with the CNN model. Performance, i.e. correctness, refers to how many models within the ensemble correctly answered each of the 32 questions. The y-axis represents the KL divergence and correctness, while the x-axis represents the documents (32 total), and the legend indicates which models the datapoints refer to. The documents on the left are part of the easier class and the divergence scales up as document difficult increases. Following the results we hypothesize, we can observe the relationship between higher performance and more similarity to humans attention (on models using traditional attention mechanisms).

Interestingly, XLNets models show a very weak correlation of -0.16 and p = 0.381 (cf. Table 2, cf. Figure 2). As most XLNet models correctly answer the questions, but divergence increases in the same scaled pattern as with the CNNs and LSTMs (cf. Figure 1, 2), we do not observe the same significant correlation between performance and similarity to human attention. We hypothesize this is may be because: the XLNet models are pre-trained on various domains, the self-attention component plays a role in these results, and/or elements within the permutation language model with varied factorization orders (Yang et al., 2019).

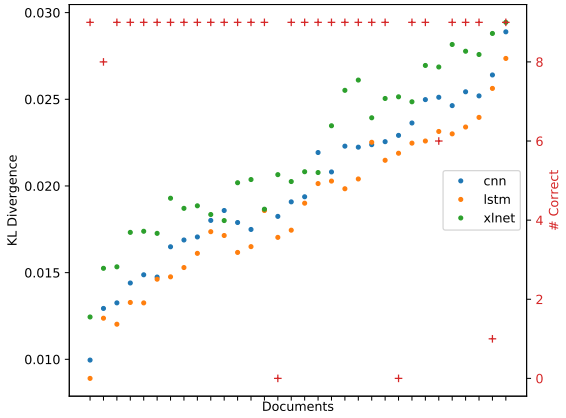

Figure 2: KL Divergence(Word Level): XLNet - All/Each Documents Ordered by the Sum of Divergence Scores and Number of Correct Models.

6.2 Analysis Results — Across Models

In Table 3, we make a pairwise comparison of the average KL divergence for the three neural models using a linear regression model with Tukey's alpha adjustment method (Sinclair et al., 2013). Interestingly, there is a **statistically significant** difference between the KL divergence of **LSTMs compared to XLNets** ($\beta = -0.003, p < 0.01$). Though the performance of the XLNets are better with respect to accuracy, LSTMs are significantly more similar to human visual attention.

When analyzing across model types, we show statistical significance that the LSTMs are more similar to humans compared to the XLNets (cf. Table 3). In addition, the LSTMs depict a trend of modeling human visual attention more than the CNN models (cf. Figure 1, 2, cf. Figure 3)

The interesting findings from our analysis on XLNet network attention when compared to the other attentive models, shows that for these transformer networks, perhaps human attention is not particularly helpful or advantageous.

Alternatively, we can see that though the XLNet outperforms the CNN and LSTM networks, achieving the newest val-set SOTA results of 91% accuracy, the KL divergence is significantly higher (from human attention) compared to the LSTM. This could show that the pre-training or self-attention method may cause this difference.

Though aiming to interpret the black box via comparison to human performance provides insight, it does not mean we should be aiming to

9-Best	Avg KL	Combo	Estimate	Std. Error	t-value	p-value
LSTM	0.018	LSTM vs. XLNet	-0.003	0.001	-2.835	< 0.01
CNN	0.020	LSTM vs. CNN	-0.001	0.001	-1.098	0.27
XLNet	0.022	CNN vs. XLNet	-0.001	0.001	-1.736	0.17

Table 3: Pairwise comparison of the average KL divergence for the three models.

force all model types to perform as human do on the same task; after all these are two very different neural systems with varying assumptions made on tasks in order to achieve high performance.

6.3 Qualitative Analysis

For the qualitative analysis, we show attention maps of the three model types and humans, over two samples from each document class (cf. Figure 3). Moreover, we show example images of our scan path visualization tool (cf. Figure 4) in which we show the human reading behavior on a portion of two sample documents from our experiments. In these results, we visualize the observed relationships uncovered in the quantitative analysis, gaining more insight on how the models and humans perform on the same comprehension task.

Furthermore, we observe that attention distribution over the easier documents (as defined by most models predicting the correct answer candidate) tend to be less diverged from human attention compared to the harder documents, and this applies across all model types. This result is interesting as it shows the relationship between performance and similarity to human attention.

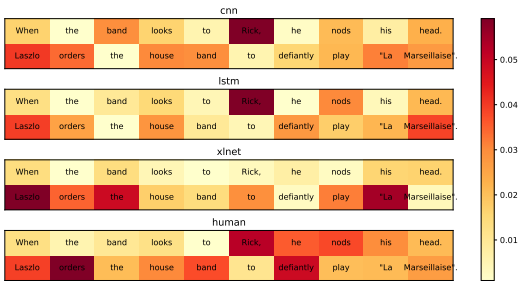

Figure 3: Example attention distributions of the three models and humans. Shown are the attention distributions over two sentences from one of the plots in the validation set.

7 Conclusion

Our core contribution is the comparative analysis between human versus various SOTA text-based attentive QA systems. To the best our knowledge, we are the first to compare human attention to neural attention, leveraging gaze data on text-based tasks. Our findings show that CNNs and LSTMs have a statistically significant negative correlation with human performance. Interestingly, the same is not true for XLNet. Moreover, the LSTM attention weights are significantly more similar to human attention compared to the XLNet. Although the pre-trained transformer networks are less similar to human visual attention, our fine-tuned model obtains the new SOTA on the MovieQA benchmark dataset with 91% accuracy on the validation set. In addition, we present our attentive reading visualization tool, to allow for qualitative analysis when comparing human versus neural attention.

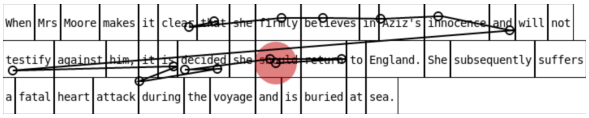

Figure 4: Example 66, scan path from our visualization tool. Shown are the reading patterns over three sentences from one of the plots in the val set.

References

David Alvarez-Melis and Tommi S Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7786–7795. Curran Associates Inc.

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

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.

Matthias Blohm, Glorianna Jagfeld, Ekta Sood, Xiang Yu, and Ngoc Thang Vu. 2018. Comparing attention-based convolutional and recurrent

neural networks: Success and limitations in machine reading comprehension. *arXiv preprint arXiv:1808.08744*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

Zihang Dai, Zhilin Yang, Yiming Yang, William W Cohen, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Heiner Deubel, K O'Regan, Ralph Radach, et al. 2000. Attention, information processing and eye movement control. *Reading as a perceptual process*, pages 355–374.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daria Dzendzik, Carl Vogel, and Qun Liu. 2017. Who framed roger rabbit? multiple choice questions answering about movie plot.

Simone Frintrop, Erich Rome, and Henrik I Christensen. 2010. Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1):6.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

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

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 2017. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258.

John M Henderson. 1992. Visual attention and eye movement control during reading and picture viewing. In *Eye movements and visual cognition*, pages 260–283. Springer.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8).

James E Hoffman and Baskaran Subramaniam. 1995. The role of visual attention in saccadic eye movements. *Perception & psychophysics*, 57(6):787–795.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329.

Ami Kristjansson. 2011. The intriguing interactive relationship between visual attention and saccadic eye movements. *The Oxford handbook of eye movements*, pages 455–470.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: Human reading inspired extractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3031–3041, Hong Kong, China. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

Yu-Fei Ma and Hong-Jiang Zhang. 2003. Contrast-based image attention analysis by using fuzzy growing. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 374–381. ACM.

Milica Milosavljevic and Moran Cerf. 2008. First attention then intention: Insights from computational neuroscience of vision. *International Journal of advertising*, 27(3):381–398.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.

Massimo Poesio. 1995. Semantic ambiguity and perceived ambiguity. *arXiv preprint cmp-lg/9505034*.

Michael I Posner. 1980. Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3–25.

Michael I Posner, Charles R Snyder, and Brian J Davidson. 1980. Attention and the detection of signals. *Journal of experimental psychology: General*, 109(2):160.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A survey on neural machine reading comprehension. *arXiv preprint arXiv:1906.03824*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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

Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506.

Raymond Reiter. 1981. On closed world data bases. In *Readings in artificial intelligence*, pages 119–140. Elsevier.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

William B Rouse and Nancy M Morris. 1986. On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3):349.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Walter Schneider and Richard M Shiffrin. 1977. Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, 84(1):1.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Richard M Shiffrin and Walter Schneider. 1977. Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. *Psychological review*, 84(2):127.

J Sinclair, Paul J Taylor, and Sarah Jane Hobbs. 2013. Alpha level adjustments for multiple dependent variable analyses and their applicability–a review. *Int J Sports Sci Eng*, 7(1):17–20.

Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.

Yaoru Sun and Robert Fisher. 2003. Object-based visual attention for computer vision. *Artificial intelligence*, 146(1):77–123.

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.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Edwin AJ Van Hooft and Marise Ph Born. 2012. Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking. *Journal of Applied Psychology*, 97(2):301.

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.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Jesse Vig. 2019. Visualizing attention in transformer-based language models. *arXiv preprint arXiv:1904.02679*.

Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *CoRR*, abs/1611.01747.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *HLT-NAACL*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

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. ACM.