論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Interpreting Reading and Writing Process of Neural Models using Eye-gaze Information
著者(和文)	Fariz Ikhwantri
Author(English)	Fariz Ikhwantri
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第12915号, 授与年月日:2024年9月20日, 学位の種別:課程博士, 審査員:徳永 健伸,岡崎 直観,村田 剛志,齋藤 豪,井上 中順
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第12915号, Conferred date:2024/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

系・コース： Department of, Graduate major in	Computer Science Artificial Intelligence 系 コース	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of (Engineering)
学生氏名： Student's Name	Fariz Ikhwantri	審査員主査： Chief Examiner	Takenobu TOKUNAGA

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Interpretation of deep neural network models is an essential topic in the natural language processing (NLP) community. Yet, the relationship between models and human behaviour in downstream tasks remains largely unexplored, calling for further investigation. Past studies have proposed various interpretation methods for neural networks, which provide saliency scores of input elements as clues to interpret a model's behaviour. On the other hand, eye movement research has a long and successful history of studying human cognitive processes. Although we cannot directly observe human cognitive processes, eye movement is believed to be a good proxy for reflecting them. Against such a background, it is natural to understand the neural network behaviour in solving NLP tasks by comparing it with the human eye-movement behaviour. This research investigates the alignment between saliency scores from neural network interpretation methods and human eye-gaze features from humans across diverse NLP tasks. Notably, this research extends beyond reading tasks like sentiment analysis, relation classification, and question answering to include writing tasks, such as summarisation.

This research aims to answer two research questions to understand the similarities and differences between models and humans in the decision process. The first question is, "Does the input word saliency from interpretation methods conform with human eye-gaze features?". The second question is, "How does the model saliency conformity impact model prediction?".

The first study is the task-specific reading. Four interpretation methods - simple gradient, integrated gradient, input-perturbation, and attention - were evaluated across three architectures: LSTM, CNN, and Transformer. Two publicly accessible corpora annotated with eye-gaze information, namely ZuCo and MQA-RC datasets, were utilised for this study. To answer the first question, I compared the models' input word saliency distance (SD) to human eye-gaze features. SD is defined as KL-divergence between the model saliency and eye-gaze feature distribution over input words. The results show that the Transformer has the highest similarity to the human gaze across reading tasks in most cases.

For the second question, I proposed a novel evaluation method called the "Saliency Distance-performance curve" (SDPC). This method visualises the cumulative model performance in relation to the SD scores. The SDPC sheds light on the underlying phenomena that were previously overlooked when solely relying on macroscopic metrics, such as average SD scores and rank correlations, as commonly done in past studies. Overall, the analysis of task-specific reading reveals that the impact of good saliency conformity between humans and machines on task performance varies based on task combinations, interpretation methods, and architectures. These findings are crucial when incorporating eye-gaze information for model training to enhance overall model performance.

In the writing task, I adopt summarisation as a target task because it involves reading a source text and writing its summary that captures the main ideas of the original text. Prior studies have analysed model interpretation in generation tasks such as translation or summarisation tasks. However, studies have yet to address how the generation process compares to that of humans. The main challenge is aligning the model saliency output in the generation process and eye-gaze features from writing activity data. The model saliency output in the generation process is a matrix form, while eye-gaze features in reading or writing tasks are vectors whose dimension corresponds to the words of the original text. I proposed a new framework for analysing summarisation models by comparing them to eye movement. The framework involves macroscopic and microscopic views of model saliency and human gaze data to handle the different representations. The model saliency output is transformed to the same representation of eye-gaze features, a vector, in the macroscopic analysis. On the other hand, eye-gaze features are converted to the same

representation of the model saliency, a matrix, in the microscopic analysis. In this study, I also built a novel dataset of extractive and abstractive summarisation by language learners, annotated with eye-gaze information and keystroke logs.

To answer the first question, I investigate the rank correlations between model saliency scores and human fixation counts in the macroscopic and microscopic analyses. Our findings suggest attention-based saliency scores partially align with human fixation counts.

For the second research question, I propose an ablation analysis which removes part of the input according to the model saliency and the eye-gaze feature. The macroscopic ablation analysis indicates that removing important words according to the human gaze can impact model performance. However, microscopic ablation analysis reveals that the human gaze does not impact model performance, which differs from macroscopic ablation. This discrepancy may be attributed to the forced decoding method, which might not accurately reflect the prediction scenario. The forced decoding uses previous ground truths, introducing bias to the decoder module. As a result, while the human gaze may influence models when using its output decoding, forced decoding can lead models to rely heavily on previous ground truths, affecting their performance.