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)	IKHWANTRIFariz
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:德永 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)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名		Fariz Ikhwantri	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	徳永 健伸	教授	審査員	井上 中順	准教授
	審査員	岡崎 直観	教授			
		村田 剛志	教授			
		齋藤 豪	准教授			

論文審査の要旨（2000 字程度）

本論文は"Interpreting Reading and Writing Process of Neural Models using Eye-gaze Information"と題し，英文 5 章と付録 4 章から構成されており，深層学習モデルがテキストの読み書きを伴う自然言語処理の課題を解く際の振舞いを同じ課題を解く際の人間の視線情報と比較分析する手法を提案し，その手法を実データに適用して分析することにより提案手法の有用性を示している．

第 1 章 "Introduction"では，本論文の背景と目的について述べている．深層学習モデルの成功は人工知能研究に大きな進展をもたらす一方で，その挙動が不明，結果の説明が困難，モデル改善の方針が立てにくいなどの問題も指摘されている．まず，モデルが課題を解く際に注目している入力要素と人間が同じ課題と解く際に注視する入力要素を比較し，深層学習モデルの挙動を解析する手法の有用性を述べている．そして，読み書きを伴う下流課題におけるモデルの挙動を分析する枠組みを提案し，実データの解析を通してその有用性を示すことが本論文の目的であるとしている．このために以下の 2 つの解明課題を立てている:(RQ1)：解釈手法が与えるモデルの注目要素と人間の注視要素は合致するか?,（RQ2)：両者の合致度はモデルの性能にどのような影響を与えるか?

第 2 章 "Related Work"では，人間がテキストを読み書きする際の眼球運動の情報を利用した研究について認知科学と自然言語処理の両方の分野で調査し，本論文との位置付けを述べている．また，近年，深層学習モデルの挙動を解釈するために，モデルが課題を解く際に入力の各要素を重視する度合い(顕現性値)を計算する様々な「解釈手法」が提案されており，その概要を説明するとともに，本論文でもそれらを使用すると述べている．

第 3 章 "Interpreting Models in Reading Tasks"では，深層学習モデルがテキストを読む際の挙動を人間の視線特徴と比較分析する枠組みを感情分析，関係抽出，質問応答の 3 つの課題について，4 つの解釈手法（単純勾配法，積分勾配法，入力摂動法，注意法)を 3 種類のモデル(LSTM，CNN，Transformer)に適用し，体系的に分析している．(RQ1)についてはモデルから得られる顕現性値と人間の視線特徴の入力要素上の分布の違いをカルバック・ライブラー情報量で定量化し，これを顕現性距離(Saliency Distance; SD)として定義している．既存のデータを用いて SD を計算し，モデルの観点からは Transformer がすべての読解課題で人間の視線によく合致し，解釈手法の観点からは短いテキストでは勾配法が，長いテキストでは注意法が人間の視線とよく合致すると報告している．さらに，(RQ2)のために顕現性距離-性能曲線（Saliency Distance-Performance Curve; SDPC)と呼ぶ新しい評価方法を提案している．SDPC は，SD 値とモデルの性能の関係を可視化するもので，先行研究で使われてきた分布間距離の平均や順位相関のような巨視的な指標では捉えにくい現象に光を当てるものである．SDPC によって既存データを分析し，モデルと人間の注目要素が合致する度合いがモデルの性能に及ぼす影響は，課題，解釈方法，モデルの種類の組合せによって変動が大き

いという知見を得ている．さらに SDPC はモデルの注目要素が人間の注視要素によく合致しているにもかかわらずモデルが誤った予測をするような特異事例を発見するのに役立つことを指摘している．

第 4 章 "Interpreting Moldels in Summarisation"では，読み書き両方を必要とする課題として要約を取り上げ，テキストの産出過程で，モデルと人間が元テキストのどの部分に注目するかを比較分析した結果を述べている．要約などの生成課題における深層学習モデルの挙動を分析した先行研究はあるが，人間の視線情報と比較して論じた研究は本論文が初めてであるとし，比較分析のための新しい枠組みを提案している．比較分析の際に考慮すべき点として，生成モデルの入力に対する顕現性分布が各単語を生成するごとに変化する離散的時系列データであるのに対し，視線特徴には各単語出力に即した明確な時系列がなく，読解課題と同様に要約完了時に得られる視線特徴の分布のみであることを指摘している．この違いを解消するために本論文では，双方の表現形式を変換する巨視的手法と微視的手法を提案している．2 つの既存データと本論文で新たに収集したデータを用いて，巨視的・微視的分析をおこない，(RQ1)については注意法による顕現性分布が人間の視線特徴分布と部分的に合致すると報告している．(RQ2)については，モデルや人間が注目する入力要素を除去してモデルに与え，その性能変化を観察する切除分析をおこなっている．巨視的切除分析では，視線特徴に従って入力を削除すると，モデルの性能に大きな影響を与えるが，微視的切除分析では，そのような影響がないという相反する結果が得られたとしている．この不一致の要因として，巨視的分析と微視的分析おける単語生成時の復号方式の違いをあげている．

第 5 章 "Conclusions"では，本論文の貢献をまとめるとともに今後の課題について述べている．

以上要するに，本論文は自然言語処理の課題を深層学習モデルが解く際の挙動を，解釈手法が与える顕現性分布として表現し，それを同じ課題を人間が解く際の視線特徴分布と比較することによって深層学習モデルの挙動を分析する手法を提案し，その手法を実データに適用してモデルの挙動を分析し，提案手法の有効性を示している．本論文の成果は深層学習モデルの分析のための新しい道具を提供し，深層学習モデルの分析のための基礎技術として位置付けることができ，工学上貢献するところが大きい．よって本論文は博士(工学)の学位論文として十分価値あるものと認める．