法政大学学術機関リポジトリ
HOSEI UNIVERSITY REPOSITORY

PDF issue: 2025-06-09

# Improving Prediction Performance and Model Interpretability through Attention Mechanisms from Basic and Applied Research Perspectives

KITADA, Shunsuke / 北田, 俊輔

(開始ページ / Start Page)
1
(終了ページ / End Page)
156
(発行年 / Year)
2023-03-24
(学位授与番号 / Degree Number)
32675甲第573号
(学位授与年月日 / Date of Granted)
2023-03-24
(学位名 / Degree Name)
博士 (工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)
(URL)
https://doi.org/10.15002/00026672

博士学位論文
論文内容の要旨および審査結果の要旨

| | |
|---|---|
| 論文題目 | Improving Prediction Performance and Model Interpretability through Attention Mechanisms from Basic and Applied Research Perspectives |
| 氏名 | 北田　俊輔 |
| 学位の種類 | 博士（工学） |
| 学位番号 | 第 573 号 |
| 学位授与年月日 | 2023 年 3 月 24 日 |
| 学位授与の要件 | 法政大学学位規則第 5 条第 1 項第 1 号該当者（甲） |
| 論文審査委員 | 主　査　彌冨　仁　　教授 |
| | 副　査　柴田　千尋　　准教授 |
| | 副　査　西田　京介　　特別研究員（NTT 人間情報研究所） |

## １．論文内容の要旨

深層学習技術に基づいた機械学習モデルは、その飛躍的な進歩に伴い、様々な分野で従来の学習モデルよりも高い性能を示すようになってきた。一方で、こうした技術に基づくモデルは具体的な予測プロセスの解釈や説明が難しいという、いわゆる Blackbox 化の問題がより顕在化してきた。予測結果の解釈性を向上させる、つまりそのモデルがなぜその結果を出力したかの説明や根拠を与える“説明可能な AI”（explainable AI）の技術がより求められるようになった。

本学位論文では、自然言語処理を行う機械学習モデルの出力の解釈性向上に関して注意機構（attention mechanisms）に着目し、導入（1 章）の後、当該分野の基礎研究（2 章、3 章）と、実応用として電子広告分野での応用研究（4 章、5 章）の両面から新しい技術の提案を行い、議論、総括を行っている（6 章、7 章）。基礎研究では、(1) 注意機構に敵対的学習(adversarial training; AT)を適用することでモデルの精度と解釈性両方を向上させる学習手法、(2) (1)の手法を半教師あり学習に拡張した学習手法の提案している。応用研究では、(3) 収益性の高い電子広告の定量的評価による作成支援、ならびに(4) 収益性が低下した電子広告の適切な打ち切り時期推定方法の提案を行っており、これらは本論文の各章を構成している。

本論文の構成は以下のとおりである。

1 章では、“説明可能な AI”について、研究の背景と歴史を具体的な文献資料を提示しながら総括するとともに、既存の機械学習モデルの解釈性における課題や、本論文で解決、議論する事柄についてまとめている。

2 章では、学習モデルの注意機構に敵対的学習（AT）を導入する新しい学習手法である Attention AT および Attention iAT （iAT: interpretable adversarial training）を提案した。提案する学習手法を、当該分野の研究論文で広く評価のために利用される構成の回帰

型ニューラルネットワーク(recurrent neural networks; RNN)に適用し、計 10 種の公開データセットを用いた文書分類、質問解答、自然言語推論タスクによる評価を行っている。その結果、提案する Attention AT/iAT は、既存の各単語の埋め込み表現に対するデータ拡張や敵対的学習よりも、単に優れた予測性能を達成するだけでなく、gradient を元にした解釈結果と高い親和性を持つ、つまりより一貫性があるモデルの高い解釈性を実現した。特に Attention iAT は、各単語の attention の強さの差を強調するように学習が行われるので、ほぼ全ての評価で最高の予測性能と、人に解釈しやすいよりメリハリのついた単語の重要度提示を実現した。また Attention AT/iAT は、従来指摘されてきた注意機構はノイズに脆弱であるという弱点を、入力文字列の各単語の注意機構に対して敵対的な学習を行うことで大幅に改善したことを報告している。

3 章では、優れた予測能と解釈性の両立を実現した Attention AT/ iAT に、仮想的敵対的学習（virtual adversarial training; VAT）を効果的に適用することで、半教師あり学習を可能にした学習手法 Attention VAT および Attention iVAT を提案している。本手法はラベルのない文書データも学習に活用できるようにすることで、モデルの予測性能と解釈性をさらに高めることができることを、前研究と同様に多くの公開データセットを用いた評価で定量的に実証した。提案する Attention VAT/ iVAT は前記学習手法と同様、注意機構を持つ機械学習モデル全てに適用できるだけでなく、学習に追加するラベルなしデータの種類を問わずに、さらなる性能向上を実現する汎用性の高さを示した。

4 章では、注意機構を活用した電子広告分野において効果の高い広告（クリエイティブ）の作成支援に関する新しい提案を行っている。提案手法は、広告テキストを入稿前に、ユーザがどの程度クリックするかの click through rate (CTR)と、クリック後に商品の購入等の行動の割合 conversion rate (CVR)を同時に予測する。提案手法では、広告配信対象や、広告のジャンルといった属性値を考慮した独自の注意機構である conditional attention 機構を提案・導入し、CTR と CVR のマルチタスク学習を行うことで、従来の予測手法や、通常の注意機構を持つ予測モデルよりも良好な予測能を、共同研究企業の実際の広告データを用いて実証した。

5 章では、電子広告の停止予測に関する、これまでにない自動化技術の提案を行っている。広告業界では効果が悪くなった広告に対して適切なタイミングでの停止が求められ、本研究はその支援を目的としている。広告停止には短期停止と長期停止の 2 種類がある業界の経験知を踏まえ、提案手法は生存時間分析を元に、ある時刻 t で停止する確率であるハザード関数の値、つまり停止確率を深層学習モデルにて推論する。学習モデルには、広告テキスト、配信対象者、広告ジャンル、これまでのクリック数や売り上げ数などの多種な情報を入力とする注意機構付きのマルチモーダル RNN を用い、テキストのどの単語が停止のために重要であるかを提示できる。このモデルの学習には、実際の業務で実際に配信された 80 万件の広告データを用いて構築され、オフライン評価に 20 万件、オンライン評価に 40 万件のデータを用いて評価を行っている。その結果、提案手法は既存の深層学習手法をベ

ースとした手法より 10 ポイント以上の差をつけて最良の結果を達成したことなど、有意な成果を報告している。

6 章では、本研究の適用範囲、本研究で議論した“解釈性”について、また後続研究に与えた影響について議論を行い、7 章で論文全体の総括を行っている。

## ２．審査結果の要旨

機械学習技術において“説明可能な AI”（explainable AI）の実現は、近年特に重要な研究課題になっている。本学位請求論文は、主に自然言語処理を行う機械学習モデルの出力の解釈性向上に関して注意機構（attention mechanisms）に着目し、基礎研究と、応用研究双方の貢献をまとめ、成果を議論するとともに、今後の展望を総括したものである。

当初、本学位論文の題名は「Attention Mechanism Meets Adversarial Perturbations: Towards Making Attention Mechanism More Robust and Interpretable as Basic and Applied Research Perspectives」であったが審査の過程で、内容をより適切に表す題名にすべきとの指摘を反映し「Improving Prediction Performance and Model Interpretability through Attention Mechanisms from Basic and Applied Research Perspectives」に変更した。

基礎研究に関しては、モデルの解釈を与える現在の主要な方法である勾配情報(gradient)に基づく解釈と、注意機構(attention)が与える解釈には一貫性がない、また既存の RNN モデルの注意機構はノイズに対して脆弱であるという Jain ら(NACCL 2019)の主張を出発点として、モデルの高い予測精度と一貫性の高いモデルの解釈双方の実現を目指した学習手法について述べている。応用研究では、大きな産業になっている電子広告分野において、売り上げ(conversion)につながる文書の入稿前評価および、売り上げが低下する広告の停止時期の予測に関する全く新しい手法の提案を行っている。それらの有効性については、共同研究企業の実際の広告配信データを用いて評価され、議論がなされている。本学位論文で述べている具体的な技術的な貢献は、基礎、応用各２つの成果に基づく下記の 4 つにまとめられ、審査の結果、これらの研究において工学的新規性と有効性を確認した。

１．学習モデルの予測精度と解釈性の両立を実現する汎用的な学習手法の提案（2 章）

学習モデルの注意機構に AT を導入する新しい学習手法である Attention AT/ iAT を提案し、一般的に相反関係にある学習モデルの精度と解釈性の両立を実現した。また注意機構を持つ全てのモデルに適用できる本提案の貢献は、ほぼすべての自然言語処理モデルの性能向上に適用できる重要な貢献である。特に Attention iAT は、注意機構の強度の差に着目して学習が行われるため、どの単語が予測に貢献するかがより明確に与えられる手法であり、従来の手法と比較して解釈性の向上が著しい。

２．１を半教師あり学習へ拡張した、精度、解釈性をより向上させる学習手法の提案（3 章）

Attention AT/iAT を、仮想的敵対的学習(Virtual adversarial training; VAT)を用いて半教師あり学習に拡張した学習手法 Attention VAT/ iVAT を提案した。これらは、教師ラベルがないデータも学習に活用し、より予測精度と解釈性を高めることができる手法であり、その効果を多くの公開データセットに対する評価実験により実証した。特筆すべきは、学習データとは関係のないラベルなしデータを学習に追加することでも、性能向上の効果が得られたことであり、言語処理モデルの性能向上において汎用的な手法となる重要な貢献である。

３．収益性の高い電子広告の定量的評価方法の提案（４章）

広告クリエイティブの効果を事前に予測する新しい枠組みを提案し、実世界の広告データを用いた評価によりその効果を実証した。技術面には、独自の conditional attention 機構を考案し、マルチタスク学習の導入によりその予測精度を向上させている。広告テキストに対して、ユーザが関心のあるフレーズを捉えて可視化が可能な attention highlighting も実装され、入稿前に重要単語の可視化に基づく、効果的な電子広告作成支援が可能となっている。この研究は実際の企業との密な共同作業に基づき行われ、ニーズや知識が十分に取り込むことができたため、極めて実践的で効果的な成果が得られた。

４．収益性が低下した電子広告の適切な打ち切り時期推定方法の提案（５章）

効果が低くなった広告に対して、適切なタイミングで運用者に停止の時期を説明可能な形で知らせることができる、これまでにない全く新しい枠組みを提案している。生存時間予測のアイディアを活用し、その予測に、注意機構を持つマルチモーダル RNN モデルを活用することで、（守秘義務のため、どの単語が停止に重要であったかの評価は行われていないが）停止予測に寄与する単語表現を解釈可能になっている。広告停止には、長期停止（摩耗）と短期停止（打ち切り）の 2 種類があるという業界の経験知を反映し、それぞれの傾向を適切にとらえるような学習モデルを提案し、それぞれ大量の実データを用い有効性を確認している。こうした電子広告停止予測はこれまでに報告されていない全く新しい研究内容であり、この成果は実用面のみならず学術面でも大きく注目に値する。

本学位論文で述べられた基礎研究の成果は後続研究で参照・活用され、応用研究の成果は、当該企業ですでに活用され実社会に貢献している。これらのことからも、申請者の研究成果は、近年特に重要視されている Explainable AI（説明可能な人工知能）の発展へに寄与しているものと認められる。本論文では、これらの成果に加え当該分野の研究動向や、現在の視点からの総括も十分に行われた。

本論文を中心に、関連する学問領域の口頭試問を行った結果からも申請者が十分な学識を

有していると判断された。よって、本審査小委員会は全会一致をもって提出論文が博士（工学）の学位に値するとの結論に達した。

（報告様式Ⅲ）