様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ４ 月 ２７ 日現在

機関番号：１４４０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８１０６

研究課題名（和文）深層学習モデルの判断根拠提示のための統一的方法の開発

研究課題名（英文）A Unified Approach for Explaining Deep Neural Networks

研究代表者
原　聡（Hara, Satoshi）

大阪大学・産業科学研究所・准教授

研究者番号：４０７８０７２１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：深層学習モデルは一般にとても複雑な構造をしており、モデルの判断根拠をユーザが窺い知ることは困難である。そのため、深層学習モデルの判断根拠を説明するための「説明法」の研究が必要となる。本研究では「統一的な判断根拠提示法」の確立に向けて「入出力間の影響度の計算」及び「類似データ・重要データの提示」の研究に取り組んだ。本研究ではこれら個別の説明法をさらに発展させるとともに、これら両者の側面を併せ持った説明法を開発した。

研究成果の学術的意義や社会的意義
深層学習モデルは高い予測・認識精度を誇る一方で、一般にとても複雑な構造をしており、モデルの判断根拠をユーザが窺い知ることは困難である。このため、深層学習モデルは一般に”ブラックボックス”とされる。”ブラックボックス”性のために深層学習モデルをそのまま人間の重要な意思決定の補助（e.g ローン審査や医療診断など）に用いることは困難である。本研究で開発した説明法はこのような深層学習モデルの”ブラックボックス”性を緩和することができる。これにより、ユーザは高精度な深層学習モデルを、その判断根拠を窺いながら意思決定補助に用いることができるようになる。

研究成果の概要（英文）：Deep neural network models are inherently complex, which hinder us from inferring the underlying mechanisms or the evidences that the models rely on when making decisions. It is therefore essential to develop "explanation methods" that can reveal such mechanism or evidence so that we can understand the decisions of the models. In this research, we focused on a unification of the popular explanation methods, the explanation by important features and the explanation by similar/relevant instances. Through the research, we deepen and improved the methodologies for each explanations individually, and we then developed a unification framework that can taken into account the advantages of the both of the explanations.

研究分野：機械学習

キーワード：機械学習　深層学習　説明可能AI

１．研究開始当初の背景

深層学習は現代の人工知能技術の根幹をなす技術である。深層学習の発展により、様々な問題（例えば画像認識や音声認識、自然言語処理など）に対して従来技術では不可能であったレベルの高い精度での予測・認識が可能となってきた。これら深層学習技術が、今後より一層私たちの日常生活へと浸透していくことに疑いの余地はない。

深層学習技術が人工知能技術に新しい地平を切り開く一方で、それら技術の社会応用には依然として大きな壁が存在する。それは深層学習モデルのブラックボックス性である[Molnar 2019]。深層学習モデルは多数の層を重ねたネットワーク構造となっており、その複雑な構造のために、入力（例えば画像）からどのような情報を読み取り、最終的な出力（例えば画像の分類結果）がどのような判断基準でもって算出されたか、を人間が理解することは非常に困難である。つまり、深層学習モデルはその出力の根拠を説明することができないのである。このようなブラックボックス性は特に人間が深く関わる分野で深層学習モデルを使う上で大きな障害となる。例えば、EU では人工知能を使って判断された結果（例えばクレジットカードの審査結果）についても、その判断について顧客への説明義務を各企業に負わせる法律が 2018 年より施行された。先に述べた通り、現在の深層学習技術ではその判断に明確な根拠を提示することができない。つまり、このような法的制約のもとでは、顧客に関する意思決定に深層学習モデルは導入できないのである。今後、人工知能技術の発展・利用の拡大に伴い、人工知能技術を安心して使えるものとして社会へ浸透させるためには深層学習モデルのブラックボックス性を解消する必要がある。

学習モデルに判断根拠を提示させる説明法は大きく分けて 2 つのアプローチがある。ここでは、p次ベクトル$x \in \mathbb{R}^p$（例えば大きさpの画像）を入力として受け取り、q次元のベクトル$y \in \mathbb{R}^q$（例えばq種類のどの画像の可能性が高いかの確率値）を出力する深層学習モデルfを考える。一つ目のアプローチは「入出力間の影響度の計算」である[Simonyan 2013; Ribeiro 2016]。これは判断に強く関連する入力要素を特定する特徴選択問題である。例えば犬の画像を認識する問題において、犬の顔部分に対応する要素x_iと出力の関連が強く、背景などの関係ない要素x_iにおける関連が弱ければ、そのモデルfはきちんと犬の顔を判断根拠に画像を認識していると言える。これらのアプローチの代表例は微分に基づく手法である[Simonyan 2013]。二つ目のアプローチは「類似データ・重要データの提示」である[Koh 2017]。これは入力と類似したデータを提示する類推問題である。犬の画像を認識する問題において、深層学習モデルが同様の犬画像を類似データとして提示できるならば、そのモデルはきちんと犬の特徴を掴んで認識していると言える。こちらのアプローチの代表例は影響関数を用いてモデルの判断と関連の強いデータを推定する手法である[Koh 2017]。

２．研究の目的

本研究の目的は、「入出力間の影響度の計算」と「類似データ・重要データの提示」という相異なる二つのアプローチを統一的な説明法を構築することである。従来、これら二つのアプローチは全く異なるメカニズムを基盤としており、そこに統一的な視点はない。本研究では新しい判断根拠提示手法の確立にあたり、特にこれら複数の異なるアプローチを統合した統一的な手法の確立を目指す。このように、従来は全く異なる技術として扱われてきたものを統一的な視点から解析・統合するところに本研究の独自性・創造性がある。

３．研究の方法

(1) 「入出力間の影響度の計算」の研究の深化

従来、「入出力間の影響度の計算」ではモデルの入力勾配が標準的に用いられてきた[Simonyan 2013]。近年の研究ではこの標準的な方法を発展させた、ユーザにとってよりわかりやすい（例えばノイズの少ないきれいな画像のヒートマップ）方法が数多く提案された。これらの研究ではモデルの構造や性質をノイズの原因と仮定し、それらの原因を取り除くことでよりわかりやすい影響度の算出方法が提案されてきた。

本研究では従来の研究のアプローチとは異なり「理想的な影響度」を定式化することを考える。そして、定式化した「理想的な影響度」を目的関数として影響度を最適化することにより「理想的な影響度」を計算する方法を構築する。

(2) 「類似データ・重要データの提示」の研究の深化

「類似データ・重要データの提示」の代表例は影響関数を用いた方法である[Koh 2017]。しかし、影響関数は対象となる問題・モデルに強い制約があり、深層学習モデルに対しては必ずしも適切な方法ではない。そのため、深層学習モデルに対して「類似データ・重要データの提示」を効果的に行うためには、強い制約を必要としないより汎用的な方法が必要となる。

本研究では従来の影響関数のアプローチを一般化し、深層学習モデルに適用できる「類似データ・重要データの提示」を開発する。特に、影響関数が課していた凸性・最適性という強い制約を取り除くことに着目する。方法としては、「深層学習モデルが確率的勾配降下法により学習さ

れる」ことを制約として用いる。これは現在の深層学習モデルの学習においては一般的な問題設定であり自然な弱い制約である。このように従来の強い制約を自然な弱い制約へと置き換えることで、深層学習モデルへと適用可能な「類似データ・重要データの提示」の方法を構築する。

(3)「入出力間の影響度の計算」と「類似データ・重要データの提示」とを統一した方法の研究

前述の２つの研究において、「入出力間の影響度の計算」と「類似データ・重要データの提示」とを十分に深化させた後に、最後にこれらを統一的に扱える方法を開発する。これら２つは全く異なる説明法であるが、人間はこれらの説明法を必要に応じて使い分ける。そのため、深層学習モデルの説明においても、必要に応じてどちらの説明法にもスイッチできるような柔軟な説明法の枠組みが必要である。

本研究では、ルールモデルをこれら２つの説明法の統合の鍵として用いる。ルールモデルはif-then-elseで記述される可読性の高いモデルである。例えば「if 体温≧38℃ then 風邪」は「体温が 38℃を超えたら風邪と判定する」ルールである。ルールモデルは「入出力間の影響度の計算」と「類似データ・重要データの提示」の両者の性質を兼ね備えたモデルである。先の体温の例では、患者の多数の特徴の中から「体温」という特徴に注目していることがわかる。これがルールモデルの「入出力間の影響度の計算」としての側面である。他方、「体温≧38℃」に合致する他のデータを収集することで「類似データ・重要データの提示」が可能となる。このように、ルールモデルは２つの説明法の橋渡し役として適した性質をもっている。そこで、このルールモデルを橋渡し役として２つの説明法を統合的に扱える新たな説明法の枠組みを構築する。

４．研究成果

(1)「入出力間の影響度の計算」の成果

「理想的な影響度」の計算を従来の「特徴選択」の問題へと帰着する方法を提案した。具体的には、「特徴選択」における前向き/後向き貪欲法のように各特徴の削除によるモデルの出力変化により各特徴の影響度を定義した。特筆すべきこととして、このようにして定義した「理想的な影響度」は入力勾配など多くの既存の「理想的な影響度」の計算方法を包含した一般的な影響度の定式化となっている。つまり、既存の「理想的な影響度」の計算方法はこの「理想的な影響度」を近似的に計算しているとみなすことができる。

本研究では、上記により定義した「理想的な影響度」を直接的に最適化する方法を提案した。また、この「理想的な影響度」を基準として、実際に計算された影響度の“良さ”を定量的に評価することを可能とした。これにより、従来よりも優れた影響度の計算法が確立され、また影響度の“良さ”の定量評価が可能となったことで、従来の方法の中にも優れた方法や大きく劣った方法があることが明らかになった。

(2)「類似データ・重要データの提示」の成果

「深層学習モデルが確率的勾配降下法により学習される」ことを制約として用いた新たな「類似データ・重要データの提示」の方法を提案した。具体的には、確率的勾配降下法の計算を遡ることで、モデルの学習の過程で重要な役割を果たしたデータを推定する方法を構築した。この方法は従来の影響関数のように凸性・最適性という強い制約を必要としないため、深層学習モデルへと自然に適用可能な方法となっている。

本手法をモデルに有害な影響を及ぼした学習データを推定する「データクレンジング」に応用したところ、適切に有害なデータを特定することができた。実際、特定された有害データを除去してモデルを再学習したところ、モデルの精度が改善することが確認できた。つまり、特定されたデータは確かにモデルの精度を低下させる有害なものであった。また、影響関数との性能比較において、提案手法の方がデータクレンジングにおいて優れた性能を有することがわかった。これは提案手法が影響関数よりも弱い制約しか要請しない、深層学習モデルにより適した手法であるためである。

(3) 統一的な説明方法の成果

「入出力間の影響度の計算」及び「類似データ・重要データの提示」の両者の側面を併せ持った説明法として、深層学習モデルを部分的に可読化する説明法を提案した。提案法では、モデルの（部分的な）可読化により、出力に影響を及ぼす重要な特徴量を説明する事が可能となった。また、可読化にルールベースのモデルを用いることで、自然に類似データを提示することも可能となった。

提案した説明法の特筆すべき点として、可読性の度合いをユーザの好みに応じて調整できる機能がある。従来の説明法は、「入出力間の影響度の計算」「類似データ・重要データの提示」のどちらにおいても、ユーザの好みとは無関係に常に画一的な説明を提供するものであった。これに対して、提案した説明法では可読性の度合いをユーザが自由に調整することで、多様な説明を提示できるようになっている。 本研究ではさらにこのような多様な説明を最適化するために、可読性と説明の精度とのトレードオフを定量化したArea Under Transparency-Accuracy Trade-off Curve (AUTAC)を提案した。AUTACが最大になるように多様な説明を最適化することで、異なるユーザが異なる可読性を要求しても常に精度の高い説明を提供できるようになった。

(4) 本研究のインパクト及び今後の展望

本研究により、「入出力間の影響度の計算」及び「類似データ・重要データの提示」という２つの代表的な説明法それぞれが発展し、さらにルールモデルを介して統一的な説明法への道筋も示された。これらの進展は、深層学習モデルの実用におけるモデルのブラックボックス性の緩和へ向けた一歩である。

また、「類似データ・重要データの提示」において「データクレンジング」における有効性が見いだせたことは特に重要な進展であった。従来、「データクレンジング」は一般にデータに関する専門知識を必要とする専門的な作業であった。これに対して、「類似データ・重要データの提示」を発展させることで、必ずしも専門知識がない開発者にも「データクレンジング」が可能であることがわかった。これは深層学習モデルの開発一般において有効な重要な技術である。

【参考文献】

[Molnar 2019] Christoph Molnar. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/, 2019.

[Simonyan 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[Ribeiro 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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, 1135-1144, 2016.

[Koh 2017] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In Proceedings of the 34th International Conference on Machine Learning, 1885-1894, 2017.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　6件）

１．発表者名 Danqing Pan, Tong Wang, Satoshi Hara
２．発表標題 Interpretable Companions for Black-Box Models
３．学会等名 The 23rd International Conference on Artificial Intelligence and Statistics（国際学会）
４．発表年 2020年

１．発表者名 Kazuto Fukuchi, Satoshi Hara, Takanori Maehara
２．発表標題 Faking Fairness via Stealthily Biased Sampling
３．学会等名 The 34th AAAI Conference on Artificial Intelligence（国際学会）
４．発表年 2020年

１．発表者名 Satoshi Hara, Atsuhi Nitanda, Takanori Maehara
２．発表標題 Data Cleansing for Models Trained with SGD
３．学会等名 Neural Information Processing Systems（国際学会）
４．発表年 2019年

１．発表者名 Satoshi Hara, Takanori Maehara
２．発表標題 Convex Hull Approximation of Nearly Optimal Lasso Solutions
３．学会等名 The 16th Pacific Rim International Conference on Artificial Intelligence（国際学会）
４．発表年 2019年

１．発表者名 Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sebastien Gambs, Satoshi Hara, Alain Tapp
２．発表標題 Fairwashing: the risk of rationalization
３．学会等名 The 36th International Conference on Machine Learning（国際学会）
４．発表年 2019年

１．発表者名 Satoshi Hara, Kouichi Ikeno, Tasuku Soma, Takanori Maehara
２．発表標題 Maximally Invariant Data Perturbation as Explanation
３．学会等名 The 2018 ICML Workshop on Human Interpretability in Machine Learning（国際学会）
４．発表年 2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
カナダ	Universite du Quebec a Montreal			
米国	University of Iowa			