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　４ 年　 ６ 月　 ３ 日現在

機関番号：２４２０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１２１２４

研究課題名（和文）テキストベースの深層学習における分類パターンの解釈支援

研究課題名（英文）Interpretation Support of Classification Rules by Text-based Deep Learning

研究代表者
砂山　渡（Sunayama, Wataru）

滋賀県立大学・工学部・教授

研究者番号：４０３１４３９８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：　本研究においては，テキストベースの深層学習において，学習された分類パターンの意味を，人間が解釈できる環境の構築を目指した．すなわち，深層学習ネットワーク，DNN（Deep Neural Network）ならびに再帰構造をもつRNN（Recurrent Neural Network）による学習結果の解釈を人間に促すインタフェースを構築した．インタフェースにおいては，学習したネットワークを可視化して，単語のラベル付きで表示を行う．
　評価実験においては，深層学習に不慣れな被験者に，インタフェース上に表示される，学習元のテキスト集合の分類パターンについての解釈を行ってもらい，その妥当性を検証した．

研究成果の学術的意義や社会的意義
　ブラックボックス問題と言われている深層学習の学習結果を解釈するための手がかりを提示するシステムを構築したこと．ならびに，単純な全結合のDNNだけでなく，再帰構造を含むRNNにより，時系列データの学習にも対応している点で学術的意義がある．
　また，深層学習で学習された知識が取り出せるようになることで，抽出した知識を人間が活用することや，コンピュータが対応できている点とできていない点を理解した上で，結果の根拠を利用できることは，結果を利用する際の納得感が異なると考えられる．また，抽出した知識を類似する他の分野に転用するなどの応用が可能となるため，社会的意義がある．

研究成果の概要（英文）：　In this study, we aimed to construct an environment in which humans can interpret the meaning of learned classification patterns in text-based deep learning. That is, we constructed an interface that facilitates human interpretation of learning results from deep learning networks, DNNs (Deep Neural Networks) and RNNs (Recurrent Neural Networks) with recursive structures. The interface visualizes the learned networks and displays them with word labels.
　In an evaluation experiment, we asked subjects unfamiliar with deep learning to interpret the classification patterns of the learned text set displayed on the interface, and verified the validity of the interpretation.

研究分野：知能情報学

キーワード：深層学習　DNN　RNN　学習結果の解釈　知識抽出

1. 研究開始当初の背景

深層学習を用いたシステムは、精度の高い出力が行える反面、その出力を与える基準が不明な場合が多い。これは深層学習によって学習されるネットワークの構造が複雑で、パラメータの数が膨大になることから、学習されたネットワークに言葉で意味を与えられないことによる。そのため、出力の根拠を与えるための「解釈性」と呼ばれる研究が進められている。

画像データを用いた学習の場合、ネットワーク中のノードを画像で表すことで意味を与えられる可能性があるが、テキストデータを用いる学習の場合は、言葉で意味を与える必要がある。テキストベースの深層学習においては、あるデータを分類した時に、どの単語に着目して分類したかを出力できる、アテンションという機構が開発されている。しかしこれは、特定のデータを分類した際の根拠を与えるもので、ネットワークの全体が、どのような基準で分類しているか、の意味づけを行えるものにはなっていない。

本課題の学術的な「問い」は、「**深層学習で学習された分類基準は何か？**」であり、この答えを人間が得られる環境を本課題で構築する。

2. 研究の目的

本研究では、テキストベースの深層学習において、学習された分類パターンを人間が理解できる環境の構築を目指す。

現在の深層学習の解釈性に関わる研究の多くは、学習の精度を重視した深層学習のネットワークモデルに対して行われるものとなっており、解釈を意識したネットワークモデルの構築や、解釈につながる情報を提示した上で、実際の解釈を支援する機能、についての研究はほとんど行われていない。

本研究により、深層学習ネットワークへの意味づけが可能になれば、学習された分類パターンを人間が知識として活用することができる。例えば、業務レポートの良し悪しを分類するネットワークであれば、そこで学習された良い業務レポートに分類される基準を、人間がレポートを書く際の指針とすることができる。

また、得られた分類基準を人間が把握できることは、深層学習が対応できるデータと対応できないデータの理解につながる。すなわち、深層学習の結果を過信しないで、深層学習を含むシステムの人間による適切な運用が期待できるようになる。

3. 研究の方法

入力テキストに、「どのような単語の組合せ、あるいは単語の出現順序が含まれていれば、どの分類先に分類されるか」の分類パターン候補を提示する仕組みを検討する。具体的には、深層学習により学習されたネットワークにおいて、出力となる分類先と強いつながりをもつ部分ネットワークを特定し、そのネットワークを分類基準となる言葉で表す。これにより、分類パターン候補を「分類基準 → 出力先」と提示する。

研究は、主に以下の３つのプロセスに基づいて実施した。

1)学習結果の解釈を可能にする深層学習モデルの開発

解釈に有効な単語を取り出すため、分類に寄与する単語、単語の組合せ、単語の出現順序を特定するための仕掛けを、深層学習モデルに導入することを検討する。特定のデータの分類に寄与

する単語を特定するアテンション機構を拡張して、分類に寄与する単語の組合せや順序を特定する機構を検討する。また、分類時に多くのデータに共通に寄与する部分ネットワークを特定する機構により、学習結果の解釈につなげる方法も検討する。

2) 学習結果の解釈支援インタフェースの開発

学習結果の適用ドメインにおける分類パターンの意味解釈を支援するインタフェースを構築する。分類パターン候補の吟味は、データ分析におけるデータの絞り込みに相当する。また、パターンへの意味づけは、データ分析における知識創発に相当する。そのため、これまでに申請者が構築している、テキストマイニングのための統合環境 TETDM（https:// tetdm.jp）において実装済みの関連機能と、それらを活用する際の知見を応用する。

3) 解釈の妥当性の検証

エンドユーザによって行われた解釈の妥当性を検証する。解釈の正しさや網羅性が高くならない場合において、その原因を 1)および 2)の方法と照らして検討し、両者の内容の改善を促す。また、解釈結果として得られた知識を、実社会でのスムーズな活用につなげるためのプロセスを検討し、活用に繋げやすい解釈の方向性を検討する。

4. 研究成果

本研究においては，テキストベースの深層学習において，学習された分類パターンの意味を，人間が解釈できる環境の構築を目指した.

2019 年度においては，ベースラインとなる解釈対象として，学習ネットワークを DNN（Deep Neural Network）とし，単語入力モデルに BoW（Bag of Words）を用いて深層学習を行ったネットワークに対して，分類基準の表現を 1 単語で表すシステムを構築し，その解釈を人間に促すためのインタフェースを構築した.

2020 年度においては，2019 年度に構築したモデルを拡張して，再帰構造をもつ RNN モデルによる深層学習ネットワークを，HMM(Hidden Marcov Model)の構造に当てはめる手法により，学習結果の解釈を促す次の機能を有するインタフェースを構築した．1)分類パターンの抽出について，学習ネットワークにおいて，分類先と強いつながりを持つパスを一定数抽出．2)抽出されたパターンを視覚的に表示するため，分類先につながるパス（ノードとリンク）を表示し，各ノードに単語を付与して，どの単語がどの分類先につながるかを明示．また，時系列的にわかりやすくなるように，ノードのラベルは，学習に利用したテキストに実際に出現する単語の時系列パターンをもとに表示.

2021 年度においては，作成したインタフェースの評価と検証を行った．評価実験においては，深層学習に不慣れな被験者が，インタフェース上に表示される，学習元のテキスト集合の分類パターンについて，特に単なる単語の組合せでなく，時系列パターンを意識したテキストの意味解釈が可能か，また，解釈した結果は妥当かを確認した．結果として，80%以上の割合で時系列を意識した解釈が行われ，またその内容のほとんどが学習元のテキストの内容に合っていたことが確認された.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 安藤雅行・河原吉伸・砂山渡・畑中裕司	４．巻 34
２．論文標題 深層学習ネットワークへのHMM適用によるテキストベースの分類パターン解釈支援	５．発行年 2022年
３．雑誌名 日本知能情報ファジィ学会誌	６．最初と最後の頁 501-510
掲載論文のDOI（デジタルオブジェクト識別子） 10.3156/jsoft.34.1_501	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Junjie Shan, Yoko Nishihara, Akira Maeda, and Ryosuke Yamanishi	４．巻 38
２．論文標題 Question Generation for Reading Comprehension Test Complying with Types of Question, Journal of Information Science and Engineering	５．発行年 2022年
３．雑誌名 Journal of Information Science and Engineering	６．最初と最後の頁 to appear
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yoko Nishihara, Seiya Tsuji, Wataru Sunayama, Ryosuke Yamanishi, and Shiho Imashiro	４．巻 14
２．論文標題 A Generation Method for the Discussion Process Model during Research Progress Using Transitions of Dialog Acts	５．発行年 2021年
３．雑誌名 International Journal On Advances in Systems and Measurements	６．最初と最後の頁 17-26
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 安藤雅行，河原吉伸，砂山渡，畑中裕司	４．巻 31
２．論文標題 テキストベースの深層学習における分類パターンの解釈支援	５．発行年 2019年
３．雑誌名 日本知能情報ファジィ学会誌	６．最初と最後の頁 779 - 787
掲載論文のDOI（デジタルオブジェクト識別子） 10.3156/jsoft.31.4_779	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計12件（うち招待講演　0件／うち国際学会　5件）

1．発表者名
Masayuki Ando, Yoshinobu Kawahara, Wataru Sunayama, and Yuji Hatanaka

2．発表標題
Interpretation Support System for Classification Patterns Using HMM in Deep Learning with Texts

3．学会等名
the 14th International Conference on Advances in Computer-Human Interactions (ACHI 2021)（国際学会）

4．発表年
2021年

1．発表者名
安藤雅行・河原吉伸・砂山渡・畑中裕司

2．発表標題
深層学習ネットワークへのHMM適用による分類パターン解釈支援

3．学会等名
第27回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会

4．発表年
2021年

1．発表者名
安藤雅行・砂山渡・畑中裕司

2．発表標題
HMMを利用した深層学習ネットワークからの分類パターンの解釈支援システム

3．学会等名
第35回人工知能学会全国大会

4．発表年
2021年

1．発表者名
Yoko Nishihara, Xinran Lin, and Ryosuke Yamanishi

2．発表標題
Do the Number of Creators and Their Conversations Affect Re-Evaluation of a Familiar Place in Making a Tourist Map?

3．学会等名
the 14th International Conference on Advances in Computer-Human Interactions (ACHI 2021)（国際学会）

4．発表年
2021年

１．発表者名
N. Takeishi, and Y. Kawahara

２．発表標題
Knowledge-Based Regularization in Generative Modeling

３．学会等名
the 29th Int'l Joint Conf. on Artificial Intelligence and the 17th Pacific Rim Int'l Conf. on Artificial Intelligence (IJCAI-PRICAI'20)（国際学会）

４．発表年
2020年

１．発表者名
Junjie Shan, Yoko Nishihara, Akira Maeda, and Ryosuke Yamanishi

２．発表標題
Extraction of Question-related Sentences for Reading Comprehension Tests via Attention Mechanism

３．学会等名
the 2020 International Conference on Technologies and Applications of Artificial Intelligence（国際学会）

４．発表年
2020年

１．発表者名
安藤雅行・河原吉伸・砂山渡・畑中裕司

２．発表標題
HMMを利用した深層学習ネットワークからの分類パターンの抽出と解釈

３．学会等名
第34回人工知能学会全国大会

４．発表年
2020年

１．発表者名
若園紫乃・砂山渡

２．発表標題
単語の組み合わせによるテキスト集合のラベルの自動生成

３．学会等名
第26回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会

４．発表年
2020年

１．発表者名
安藤雅行，河原吉伸，砂山渡，畑中裕司

２．発表標題
HMMを利用した深層学習ネットワークからの分類パターンの抽出と可視化

３．学会等名
第23回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会

４．発表年
2019年

１．発表者名
安藤雅行，河原吉伸，砂山渡，畑中裕司

２．発表標題
深層学習における学習ネットワークからの分類パターンの抽出

３．学会等名
第33回人工知能学会全国大会

４．発表年
2019年

１．発表者名
若宮悠希，砂山渡，畑中裕司，小郷原一智

２．発表標題
深層学習を用いたTwitterユーザからの趣味情報の抽出

３．学会等名
第24回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会

４．発表年
2020年

１．発表者名
Junjie Shan, Yoko Nishihara, Ryosuke Yamanishi, and Akira Maeda

２．発表標題
Question Generation for Reading Comprehension of Language Learning Test

３．学会等名
2019 International Conference on Technologies and Applications of Artificial Intelligence（国際学会）

４．発表年
2019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
研究分担者	河原　吉伸 (Kawahara Yoshinobu) (00514796)	九州大学・マス・フォア・インダストリ研究所・教授 (17102)	
研究分担者	西原　陽子 (Nishihara Yoko) (70512101)	立命館大学・情報理工学部・教授 (34315)	

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