様式　10

博士学位論文審査報告書

申請者氏名（ふりがな）　安藤 雅行（あんどう まさゆき）

学位の種類　博士（工学）

論文題目　深層学習の重みネットワークを用いた
テキスト分類パターンの解釈支援

学籍番号　1968001

学歴　平成 29 年 4 月　滋賀県立大学大学院工学研究科
電子システム工学専攻博士前期課程入学
平成 31 年 3 月　同上修了
平成 31 年 4 月　同研究科
先端工学専攻博士後期課程進学
令和 4 年 3 月　同上~~修了~~修了見込

論文審査委員　（主査）滋賀県立大学工学研究科　教授　砂山 渡　印
滋賀県立大学工学研究科　教授　酒井 道　印
滋賀県立大学工学研究科　教授　南川 久人　印
九州大学マス・フォア・インダストリアル研究所
教授　河原 吉伸　印
大分大学理工学部　教授　畑中 裕司　印

論文の内容の要旨

本論文は，日本語コーパスを学習させた深層学習モデルを対象として，学習結果の根拠を示す分類パターンを，学習によって構築された重み付きネットワークから抽出して，その分類パターンの解釈を支援するシステムの構築と，その有効性の検証について述べている．

本論文は 5 章から構成されている．

第 1 章では，本研究の背景と目的，課題に対するアプローチなどについて述

べている．まず，深層学習がどのように発展したのか，どのような種類があるか，その過程で発生した問題は何かなど，深層学習の歴史について解説している．そして，近年増加する深層学習の活用例や，今後どのように使われていくかなどの最新研究について述べた後，現在，深層学習において重要視されている問題点として，深層学習のブラックボックス問題を取り上げている．その上で，ブラックボックス問題に対する本研究のアプローチとして，自然言語処理分野での，深層学習の重みネットワークを用いたテキスト分類パターンの解釈支援システムについて述べている．その後，論文全体の構成について述べている．

第２章では，本研究に関連する研究として，まず，深層学習の出力に寄与する入力に注目した研究について，特にアテンションと呼ばれる手法について先行研究を紹介している．また，本研究との違いとして，アテンションは学習済みネットワークの内部に注目しているわけではないという点を述べている．次に，深層学習の学習ネットワークの解釈に注目した研究について，画像処理分野で中間層の学習過程に注目した先行研究について述べている．また，本研究との違いとして，画像処理分野では中間層の画像化が容易であるが，自然言語処理分野では中間層の可視化が難しい点を述べている．最後に，深層学習の説明可能性に注目した研究について，説明可能なAIについての先行研究を紹介している．また，本研究との違いとして，本研究では，モデルの精度や信頼性などの評価ではなく，人間が学習結果に納得するためのシステムの構築を行っていることについて述べている．

第３章では，単純な深層学習モデルである，DNN（Deep Neural Network）から分類パターンを抽出し，解釈支援を行うシステムの構築について述べている．ここでは，学習済みの DNN の重みの値によって情報の重要度が決まるとし，重みの積から出力に寄与する特徴，または，各中間層ノードに寄与する特徴を取得する手法を提案している．また，取得した情報は可視化インタフェースによって解釈支援システムの利用者に提供していることを述べている．解釈支援システムの評価実験として，動物の生態に関する文章や映画のレビュー，受験に関するツイート集合などを題材とし，14 人の被験者に提案システムを用いて解釈を行ってもらったことを述べている．得られた解釈の内容が学習に用いた文章集合に対して妥当かどうか調査したところ，90%近くの解釈が妥当であったとして，提案システムの有効性を確認している．

第４章では，実際に自然言語処理分野で広く扱われている RNN（Recurrent Neural Network）モデルについて，分類パターンを抽出し，解釈支援を行うシステムの構築について述べている．ここでは，RNN の重み付きネットワークが HMM（Hidden Markov Model）と類似している点に注目して，RNN の重みを一種の

確率変数をみなして一つの HMM と捉えることを述べ，単語の時系列パターンに対する尤度を算出することで，時系列情報を含んだ分類パターンを取得する手法を提案している．また，取得した分類パターンは，単語をノード，単語の時系列を矢印付きエッジとする解釈支援ネットワークとして表示し，解釈支援システムの利用者に提供することを述べている．解釈支援システムの評価実験として，アニメの登場人物のセリフ集合や Amazon の家電商品やゲームソフトのレビュー集合を題材とし，8 人の被験者に提案システムを用いて各文章集合に対する解釈を行ってもらったことを述べている．提案システムにおいて解釈内容が文章集合に対して妥当かどうか調査したところ，95%近くの解釈が妥当であったとして，提案システムの有効性を確認している．

第 5 章は結言で，本研究において，日本語コーパスを学習させた深層学習ネットワークにおける，分類根拠を表す分類パターンの解釈を支援するシステムの構築を，DNN と RNN の 2 つの深層学習モデルを対象として行ったこと述べている．また，構築したシステムの有効性を検証する実験として，DNN のシステムにおいては，分類パターンに対する被験者の解釈の 90%近くが妥当であると確認され，RNN のシステムにおいては，分類パターンに対する被験者の解釈の 95%近くが妥当であると確認されたことから，両解釈支援システムは一定の効果があったと結論を述べている．

論文の審査結果の要旨

本論文では，日本語コーパスを学習させた深層学習モデルを対象として，学習結果の根拠を示す分類パターンを，学習によって構築された重み付きネットワークから抽出して，その分類パターンの解釈を支援するシステムの構築と，その有効性の検証についてまとめた．

第１章では，近年の深層学習が発展してきた経緯，および深層学習の応用事例について述べた上で，深層学習において重要視されている問題点として，深層学習のブラックボックス問題を取り上げ，自然言語処理分野における，深層学習の重みネットワークを用いたテキスト分類パターンの解釈支援システムの必要性について論じた．

第２章では，関連研究として，深層学習の出力に寄与する入力に注目するアテンションと呼ばれる手法を用いた研究，画像処理分野で深層学習の学習ネットワークの中間層の学習過程に着目した研究，ならびに深層学習の説明可能性に関する研究について論じた．

第３章では，最も基本的な深層学習モデルの DNN (Deep Neural Network) を用いたテキスト分類問題を対象として，深層学習ネットワークで学習された重み情報をもとに，分類パターンを表す単語の組合せを可視化するシステムを提案し，人間が学習結果の解釈を行うことを支援するシステムを提案した．被験者が，同システムを用いて学習結果の解釈が可能か否かを検証する実験を行い，被験者による解釈の約 90%が妥当であったことを確認した．

第４章では，時系列データを扱う深層学習モデルの RNN (Recurrent Neural Network) を用いたテキスト分類問題を対象として，深層学習ネットワークで学習された重み情報を，HMM (Hidden Markov Model) によるネットワークとして表現し，分類パターンを表す単語の出現順序を可視化するシステムを提案し，人間が学習結果の解釈を行うことを支援するシステムを提案した．被験者が，同システムを用いて学習結果の解釈が可能か否かを検証する実験を行い，被験者による解釈の約 95%が妥当であったことを確認した．

第５章では，テキスト分類問題を対象として，DNN と RNN の２つの深層学習モデルを用いて，深層学習による分類根拠を表す分類パターンの解釈を支援するシステムを構築し，その有効性を確認したことを記した．

本論文に示された成果は，学術的視点においては，深層学習の学習結果の説明可能性を高めるための研究として，技術的観点と人間の解釈を支援するインタフェースの観点の両面で多くの可能性を示唆するものとなっている．また，工学的視点においては，テキストデータからの知識の抽出と伝承に関わる内容として，データからの知識発見やスキル伝達など幅広い応用の可能性を有して

いる．そして，以上の研究業績は，工学研究科における課程修了による博士の学位授与に関する内規の運用方針に定める基準が示す要件を満たしている．

以上に基づいて，本論文は博士（工学）の学位論文として価値があるものと認める．また令和3年12月21日の公聴会に引き続き実施した最終試験の結果，合格と判定した．