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(第21回)
SIG-AM-21-09

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

Extraction of Classification Patterns from Deep Learning Networks

安藤 雅行[1*]　河原 吉伸[2,3]　砂山 渡[4]　畑中 裕司[4]
Masayuki ANDO[1]　Yoshinobu KAWAHARA[2,3]　Wataru SUNAYAMA[4]　Yuji HATANAKA[4]

[1] 滋賀県立大学大学院工学研究科
[1] Graduate School of Engineering, The University of Shiga Prefecture
[2] 大阪大学産業科学研究所
[2] The Institute of Scientific and Industrial Research
[3] 理化学研究所革新知能統合研究センター
[3] RIKEN Center for Advanced Intelligence Project
[4] 滋賀県立大学工学部
[4] School of Engineering, The University of Shiga Prefecture

Abstract: In deep learning, there is a problem that concrete classification patterns for deriving reasons for classification are often incomprehensible. In this paper, we propose a classification patterns extraction system from deep learning networks and verified the effectiveness of the system. The proposed system takes out learning networks from the learning result of deep learning and extracts classification patterns from the learning networks. Then the system displays the extracted classification patterns so that users of the system can interpret the learning networks. In verification experiments, the significance of the extracted classification patterns was estimated by chi-square test. The results showed that users of the system can extract classification patterns effective for interpretations of the learning networks by using the proposed system.

1　はじめに

インターネットの普及に伴い，また，SNS（Social Networking Service）の出現によって，画像，テキスト，数値データが大規模になり，その処理や情報の抽出に機械学習が使用されるようになってきた．しかし，従来の機械学習は大量のデータから規則などを学習し，分類・予測を行う際，データのどの特徴（画像なら色や形など）に注目するかは人間が指定する必要があった．そこで注目されるようになってきた技術が，深層学習である．深層学習は近年流行りだした機械学習であり，学習を行う層（入力データの規則などを学習する部分）を多層化している．これにより，より人間の脳の学習に近い段階的な学習ができ，従来の機械学習と比べて学習の精度が高いという利点がある．

一方で，その深層学習による予測・分類基準が人間には不明な点が問題になってきている．特に，医療分野や自動運転では，その分類基準の理解は安全性において重要視されている．仮にテキスト分野においても深層学習の判断基準をより深く理解できれば，医療分野において新人とベテランの書いた電子カルテの違いから，良い電子カルテを書く方法を容易に理解でき，企業においても良い報告書や企画書を書く方法を短時間で習得できるなど，深層学習の新しい活用が期待される．

本研究では，構造が複雑になる代わりに，単語の出現の時系列や順序も考慮した学習が可能な，再帰的深層学習（Recurrent Neural Network や Long short-term memory など）を使用し，テキスト集合の学習によってネットワークの層に付けられた重みの値を取り出し，各層を構成するノードと呼ばれる要素が持つ情報を，単語として表現する．そして層間の単語の結びつきに時間の流れ（単語のテキスト中での出現順序）を当てはめることで，単語の順序を考慮した組み合わせとしての分類のパターンの抽出を行うシステムを提案する．

以下本論文では，2 章で関連研究について述べる．3.2 章で深層学習による分類パターンの解釈支援システムの構成と詳細について述べる．4 章で提案システムの評価実験について述べ，5 章で本論文を締めくくる．

*連絡先：滋賀県立大学大学院工学研究科 電子システム工学専攻 安藤雅行
〒 522-8533 滋賀県彦根市八坂町 2500
E-mail: oh23mandou@ec.usp.ac.jp

2 関連研究

インターネットの普及などにより，急速に大規模化しつつあるテキストへの対策として活用され始めているのが，深層学習を用いたテキストマイニングシステムである [1, 2]．深層学習とは，一般に多層から構成されるニューラルネットワークを用いた学習を指し，例えば，深層学習の応用モデルである畳み込みニューラルネットワーク [3] の出現により，画像を用いた場合に限らず多くの場面で高い分類性能を実現できることが報告されている．

その一方で，深層学習は，その出力を導いた根拠についての解釈が困難であることも知られている．画像認識においては，この問題に対する研究も最近進められており，例えば，入力画像に対応する畳み込みニューラルネットにおける層間のスコアの勾配を計算することでネットワークの可視化を行う方法 [4] や，学習済みのネットワーク中間層のノード情報を用いて，対応する画像中の画素への寄与度を計算することにより画像の分類に重要な部位を表示する方法などが提案されている [5]．

しかし自然言語への深層学習の適用においては，上記のような画像認識における方法を直接適用できない．そこで，アテンションと呼ばれる手法を用いた研究 [6, 7] が注目されている．アテンションとは，深層学習において分類・予測を行う際，出力に直接結びつく入力を探索する手法で，このアテンションにより，出力に貢献する特徴は何かが視覚的にわかりやすくなっている．最新の研究では，アテンション計算を層ごとに行い，より分類・予測精度を高めた研究 [8] や，アテンションのみで構築された深層学習 [9] なども登場している．しかし，アテンションはあくまで入力と出力の関係のみに注目し，内部でどのような学習が行われているかは考慮していない．

そこで，自身の学士の研究 [10] では，テキストベースの深層学習について，層ごとの学習の流れを単語情報として表し，人間が理解できる形に直すことで，分類基準の理解のための，学習ネットワークの解釈を支援するシステムの開発を目的とし，一定の成果を得ることができた．一方で，この時使用した深層学習が，構造は単純だがテキストの単語の有無だけを特徴とし，単語の出現の時系列や順序を一切考慮しないものだったため，学習ネットワークの解釈が一定までしか得られなかった．

したがって，本研究ではこのような問題意識の下，文章（テキスト）の分類問題を例として，時系列関係を含めた分類に寄与する出力ごとの特徴を抽出できるように，再帰的ニューラルネットワークを用いた学習によってネットワークの各層に付けられた重みの値を抽出し，タイムステップごとの中間層が学習した情報を

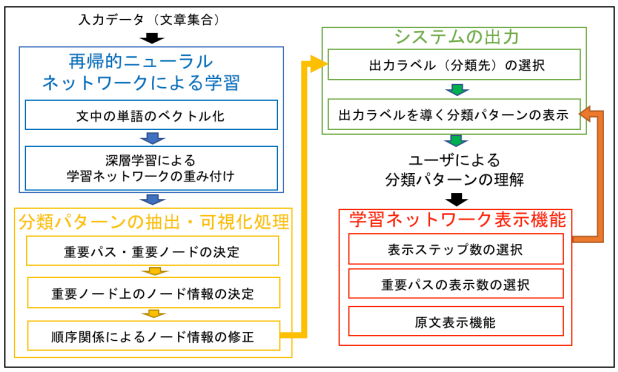

図 1: 分類パターン抽出システムの構成

文章中の単語として表現する．そこから，単語の順序関係を考慮した組み合わせとしての分類のパターン，つまり出力を導くルールの抽出を行うシステムの構築を目指す．

3 深層学習の重みを用いたテキストの分類パターン抽出システム

本章では，本研究で開発した深層学習の重みを用いたテキストの分類パターン抽出システムについて，システムの構成とその詳細について述べる．

3.1 分類パターン抽出システムの構成

分類パターン抽出システムでは，まず，図 1 に示すように，各分類先ごとにラベル付けしたテキスト集合を RNN にて分類し，その分類先を導いた学習ネットワークと，学習ネットワーク上の重みから，提案システムの分類パターンの抽出・可視化処理によって各出力 (分類先) を導くネットワーク上のパスと，パス上の各ノードの情報（そのノードで学習された単語）の決定，表示を行う．最後に，システムの利用者は，システムによって得られた学習ネットワークの表示を自分が見やすいように調整し．分類パターンを抽出する．そして分類パターンの意味を理解しやすくするための機能を利用できる．

3.2 深層学習による学習ネットワークの形成

3.2.1 文中の単語のベクトル化

深層学習で学習を行う前に，テキストデータは文中の単語を抽出したあと，単語は One hot 法 [11] と呼ばれる手法に従い単語ベクトルの羅列に直される．そして，文中の各単語を単語ベクトルに置き換え，深層学

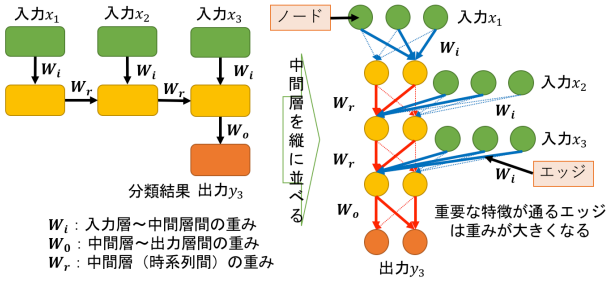

図 2: RNN の学習ネットワークと学習の様子

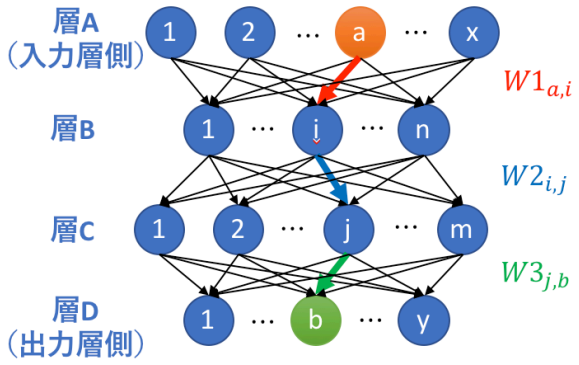

図 3: 4 層全結合型ネットワーク

習への入力データとする．なお，抽出される単語は名詞する．これは，文章の特徴とその順序関係をより学習・抽出しやすくするためである．

3.2.2 学習によるネットワークの重み付け

One hot 法によって単語ベクトルの羅列に変換され，さらにラベル付けされたテキストデータは，RNN の学習にて，それぞれの出力ラベル（分類先）を導くネットワークへの重み付けがされていく．その様子を図 2 に示す．入力文章は各単語がベクトル化され，タイムステップごとに単語ベクトルが順番に入力されていく．また，RNN での分類時は，最後の単語が入力されたタイミングで，出力層から出力される．

3.3 学習ネットワークからの分類パターンの抽出・可視化処理

本提案システムの分類パターンの抽出・可視化処理では，RNN によって得られた学習ネットワークから，各出力を導く最も関係が強いパス（ネットワーク上のエッジの繋がりの線）を決定する処理を行う．この出力と繋がりが強いパス（重要パスと呼ぶ）の決定について，図 3 に示した 4 層全結合型ネットワークモデルを例として，具体的な手順を述べる．

まず，ある分類先（図 3 の例では層 D のノード b）に到達するパスについて，パス上のエッジについた重みの積で定義される，重要度と呼ばれる値を算出する．図 3 の入力層のノード i からの太矢印のパス上のエッジに付いた重みを $W1_{a,i}$，$W2_{i,j}$，$W3_{j,b}$ とすると，このパスの重要度 $P_{a,i,j,b}$ は以下の式 (1) で導かれる．

$$P_{a,i,j,b} = W1_{a,i} \times W2_{i,j} \times W3_{j,b} \tag{1}$$

そして，式 1 で出力 b に到達する全てのパスの重要度を計算して比較し，最も値の大きいパスを，重要パスと決定する．出力 b の重要パスの重要度を S_b として，その計算式 (2) を以下に示す．図 2 の右図のように，出力層と中間層間の重み $\boldsymbol{W_o}$ のあと，中間層間の重み $\boldsymbol{W_r}$ を繰り返し辿ることで，過去の中間層を遡ることになる．

$$S_b = \max_{i,j,b} B_{a,i,j,b} \tag{2}$$

次に，重要パス上のノードについて，ノードの情報を決定の処理を行う．ノード情報の決定方法は，出力ごとの重要パスを決定した時と同様に，入力層からノード情報を決定したい中間層ノード間の重み $\boldsymbol{W_i}$ の大きさから，ノード間の結びつきの強さを求める．ただし，図 2 に示す様に，各中間層にはそれぞれ個別の入力層が対応している．ここで，RNN の入力には One hot 法による単語ベクトルを用いているため，入力層ノードにはノードごとに 1 種類の単語が対応していることになる．そこで，重要パス上の入力層ノードの単語を参照し，その単語をノードの情報と決める．最後に，ノード情報の単語について，前後の中間層ノード情報の単語と，原文中でその順番で表示されているかどうかを確認し，されていないなら単語の重要度を下げる．

こうして各出力を導く重要パスとパス上のノード情報を表示した学習ネットワークから，過去の中間層の重要ノードから現在の中間層の重要ノードの単語より，時系列を考慮した特徴の並びとしての分類パターンを抽出することができる．

3.3.1 出力ラベルを導く分類パターンの表示

本研究で開発した分類パターンの抽出システムでは，分類先に強く結びつく，重要パスの集合としての学習ネットワーク上に重要ノードの情報が表示される．例として，5 種類のお菓子の作り方に関するテキストの分類を行った場合の，システムのメイン画面を図 4 に示す．表示分類先は「マカロン」とする．このネットワークは，RNN が学習した情報を，過去から順番に中間層上に表示したものである．過去から出力層直前ま

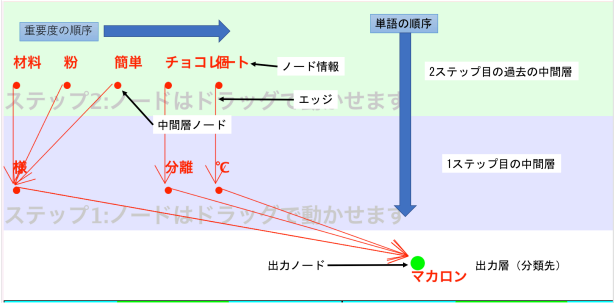

図 4: システムの画面

表 1: 学習ネットワーク表示機能

機能名	効果
表示するステップ数の選択	RNN において，何ステップ過去までの中間層を表示するか決定する
重要パスの表示数の選択	分類先ごとに，何本の重要パスを表示するか決定する
原文表示機能	分類パターン（順序を考慮した単語の組み合わせ）が，原文中でどのように出現しているかを表示する

での中間層の重要ノード情報（単語）の並びは，分類パターンを示し，重要パス 1 本につき 1 つの分類パターンが表示されていることになる

図 4 では，入力層を除いて選択したステップ数（図 4 では 2）だけ中間層と，出力層が表示され，出力層には分類された分類先を示す出力ノードが表示されている．また，分類先を示す出力ノードには，そこから選択したステップ数の長さで，選択した重要パス数だけ重要パスが表示されている．そして，重要パス上の各中間層ノードには，ノード情報として単語が表示されている．

3.4 システム上での学習ネットワーク表示機能

システムには，利用者が分類パターンの抽出・理解を行いやすいように，その表示内容を変更できる機能がある．その主なものを表 2 に示す．

4 分類パターン抽出システムの有効性の検証実験

本章では，深層学習の重みを用いた分類パターンの抽出システムにより，学習ネットワークから抽出できる分類パターンが，テキストデータを理解する（学習ネットワークの解釈を行う）を目的とした上で，各分類先に特徴的な分類パターンが抽出できているかを検証した実験について述べる．なお，本実験では，分類パターンを「2 種類の異なる単語の，順序を考慮した組み合わせ」と定義する．単語を 2 種類にした理由は，最も基本的な分類パターンであるためである．

表 2: テキストデータの詳細

データ名	内容
童話 5 種類	日本の童話「かぐや姫」「鶴の恩返し」「さるかに合戦」「桃太郎」「浦島太郎」の，それぞれの概要やあらすじなどについて書かれたテキストをネット上[1]から 1 種類あたり 50 テキストずつ用意した．

表 3: 有意性のある分類パターンの総数

分類先名	カイ二乗検定での分類パターン総数	分類パターン総数
かぐや姫	93,211	15,527,540
鶴の恩返し	9,208	15,527,540
さるかに合戦	24,839	15,527,540
桃太郎	46,749	15,527,540
浦島太郎	53,334	15,527,540
平均	45,468	15,527,540

4.1 実験準備

4.1.1 使用テキストデータと学習モデル

分類パターン抽出の対象とするテキスト「童話 5 種類」について，詳細を表 2 に示す．また，計算上の分類パターンの総数（抽出単語数 × 抽出単語数 − 抽出単語数）とその中で「ある分類パターンが，ある分類先に特有のものである」という仮説を有意水準 5%で検定したカイ二乗検定で，有意性があると推定された分類パターン数を表 3 に示す．

続いて，深層学習として使用した RNN モデルの概要を表 4 に示す．システム上での表示設定は，学習ネットワーク表示機能により，表示ステップ数 2，重要パス数 20 とする．また，テキスト「童話 5 種類」に対する RNN の分類精度は 100%であった．

4.2 実験手順

実験の対象者は著者 1 名とした．対象者はテキスト「童話 5 種類」について，本研究で構築した分類パター

[1]Google：https://www.google.co.jp で「童話名　あらすじ」と検索し，表示された上位 50 位のサイトの本文をテキストとして利用

表 4: 学習モデルの詳細

データ名	抽出した単語集合	入力層ノード数	中間層ノード数	出力層ノード数
童話 5 種類	名詞 3,941 種の単語	3,941	50	5

表 5: 抽出された分類パターン

順位	かぐや姫	鶴の恩返し	さるかに合戦	桃太郎	浦島太郎
1 位	人→命	家→鶴	自分→一緒	腰→日本一	心→者
2 位	前→一つ	日→子供	柿→甲羅	中→男	家→者
3 位	竹→一つ	粗末→金	蟹→成長	人→日本一	声→事
4 位	年→一つ	昔→彼	臼→外	退治→日本一	魚→事
5 位	男→地	日→者	家→外	中→日本一	二→誰か
6 位	人→人々	家→不思議	芽→甲羅	成長→達	乙姫→浜
7 位	者→一つ	前→子供	木→仇	桃→日本一	村→心
8 位	姫→幸せ	心→帰り	蜂→仇	日→日本一	的→奥
9 位	頃→一つ	矢→金	的→物語	桃→幸せ	玉手箱→煙
10 位	始まり→駄目	三→百	自分→上	山→人	浦島→心

表 6: カイ二乗検定による分類パターン

順位	かぐや姫	鶴の恩返し	さるかに合戦	桃太郎	浦島太郎
1 位	竹→姫	鶴→娘	種→柿	桃太郎→退治	浦島→太郎
2 位	竹→月	鶴→家	柿→種	桃太郎→鬼	浦島→玉手箱
3 位	姫→月	鶴→一	種→木	鬼→桃太郎	太郎→玉手箱
4 位	竹→中	羽→娘	種→臼	鬼→退治	浦島→中
5 位	月→姫	一→晩	種→実	退治→鬼	浦島→乙姫
6 位	中→姫	晩→娘	柿→栗	桃→鬼	浦島→人
7 位	竹→人	鶴→晩	木→柿	退治→桃太郎	太郎→中
8 位	姫→人	一→羽	種→栗	桃→桃太郎	太郎→乙姫
9 位	中→月	娘→一	柿→臼	桃→退治	乙姫→玉手箱
10 位	人→姫	日→羽	種→家	川→鬼	浦島→日

ン抽出システムによって，各分類先ごとに 20 個分類パターンを抽出した．そして抽出された分類パターンとカイ二乗検定で有意性があると推定された分類パターン上位 20 個（カイ二乗値順）とを比較し，抽出された分類パターンが他の童話にはない各童話に特有の時系列パターンとなっているかを検証した．

4.3 結果と考察

抽出された分類パターンのうち，重要度順に上位 10 個を表 5 に示す．また，比較対象として，カイ二乗検定で有意性があると推定された分類パターンのうち上位 10 個を表 6 に示す．

まず，表 5 をみると，「かぐや姫」では**人→命**（4 人のうち一人が命を失った）や**竹→一つ**（竹のう一つが光っていた）など，「鶴の恩返し」では**家→鶴**（家に人に化けた鶴が来た）や**粗末→金**（機を売って粗末な暮らしに金が入ってきた）など，「さるかに合戦」では**柿→甲羅**（投げた柿が蟹の甲羅に当たった）や**蟹→成長**

(子蟹が成長した）など,「桃太郎」では**腰→日本一**（腰に日本一のきびだんごを付けていた）や**中→男**（桃の中に男の子がいた）など，そして「浦島太郎」では**心→者**（心の優しい若者がいた）や**乙姫→浜**（乙姫に別れを告げて元の浜に帰ってきた）など，それぞれの物語特有の特徴を持つ分類パターン（赤字で表記）が平均 7 個ほどと，多く抽出できたことがわかる．よって，本研究の提案システムでは，利用者が分類先に特有の分類パターンを抽出することができるようになると言える．

一方で，表 6 をみると，カイ二乗検定での分類パターンは,「かぐや姫」では**竹→姫**,「鶴の恩返し」では**鶴→娘**など，物語の大筋に関係がありそうな単語を含むものが多く，こちらの方が学習ネットワークの解釈に有効そうに見える．しかし，各分類先ごとの分類パターンには，意味が重複している余計な分類パターン（青字で表記）が上位に来ており，それらは学習ネットワークの解釈においては余分と考えられる．そして，カイ二乗値だけではどれが余分でどれが重要な分類パターンかを判別することは難しい．よって，学習ネットワークの解釈という点では，提案システムによる抽出された分類パターンの方が有効的と言える．

以上より，実験結果から，本研究で開発したシステムでは，カイ二乗検定での分類パターンと比較することで，深層学習の学習ネットワークから，学習ネットワークの解釈に役立つと思われる，分類先に対して特有の特徴を表す，有効性のある分類パターンの抽出が可能であると言える．

5 おわりに

本研究では，複数のテキストデータの分類を，単語の順序関係を学習できる深層学習である RNN で行い，学習ネットワークの解釈を行うための，分類パターンの抽出システムの構築を目的とした．本研究の特徴として，重みを辿ることで，RNN の学習ネットワーク内の情報の伝達を過去に向かって探索している点が挙げられる．提案システムの有効性を確かめる検証実験では，提案システムで抽出された分類パターンとカイ二乗検定で有意性があると推定された分類パターンを比較し，提案システムによって抽出された分類パターンの方が，テキストの特有の特徴を理解するのに有効的であると結論づけた．今後の研究では，抽出された分類パターンからテキスト自体への解釈を行えるよう，分類パターン同士の組み合わせを用いた学習ネットワークの解釈支援を目標とする．

参考文献

[1] ボレガラ ダヌシカ，“自然言語処理のための深層学習”，人工知能学会誌，Vol.29, No.2, pp.195–201, 2014

[2] Ebru Arisoy，Tare N. Sainath，Brian Kingsbury，Bhuvaba Ramabhadran，“Deep Neural Network Language Models”, In Proceedings of the NAACLHLT Workshop，Will We Ever Really Replace the N-gram Model?, pp.20-28, 2012

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, In Proceedings of the IEEE, 1998

[4] Matthew D. Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks”, In Proceedings of ECCV ’14, pp.818-833, 2014

[5] 西銘 大喜，“ディープニューラルネットワークによる画像からの表情表現の学習”，第 29 回人工知能学会全国大会，3L4-3，2015

[6] M Daniluk, T Rocktaschel, J Welbl, S Riedel, “Frustratingly Short Attention Spans in Neural Language”, ICLR, 2017

[7] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need”, CoRR, vol. abs/1706.03762, 2017

[8] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. “Convolutional sequence to sequence learning”, arXiv preprint arXiv:1705.03122v2, 2017

[9] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N.Gomez, L.Kaiser, I.Polosukhin, “Attention Is All You Need”, In the Annual Conference on Neural Information Processing Systems (NIPS), 2017

[10] 安藤雅行，河原吉伸，砂山渡，畑中裕司，“深層学習における分類パターンの解釈支援”，2018 年人工知能学会合同研究会 SIG-AM-20-02，pp.1-6，2018

[11] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter.“Efficient and robust automated machine learning”, In Neural Information Processing Systems (NIPS), 2015