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27回)
SIG-AM-27-01

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

Interpretation Support System for Classification Patterns from Deep Learning Networks using HMM

安藤 雅行[1*]　河原 吉伸[2,3]　砂山 渡[4]　畑中 裕司[5]
Masayuki ANDO[1]　Yoshinobu KAWAHARA[2,3]　Wataru SUNAYAMA[4]　Yuji HATANAKA[5]

[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] 理化学研究所革新知能統合研究センター
[2] RIKEN Center for Advanced Intelligence Project
[3] 九州大学 マス・フォア・インダストリ研究所
[3] Institute of Mathematics for Industry, Kyushu University
[4] 滋賀県立大学工学部
[4] School of Engineering, The University of Shiga Prefecture
[5] 大分大学理工学部
[5] Faculty of Science and Technology, Oita University

Abstract: This paper describes an interpretation support system for classification patterns extracted from deep learning with texts using HMM, and verified its effectiveness. It is well known that classification patterns by deep learning models are often difficult to interpret the reasons derived. In the proposed system, the content of deep learning results is extracted using HMMs, and classification patterns are provided for the system users to interpret the learned features. In verification experiments to confirm the effectiveness of the system, based on the learning result of deep learning classifying sentences, In the experiment, one group used the proposed system. The other group used the system that displays words with high TFIDF values. The results show that the subjects who used the proposed system were able to understand the meanings of the classification patterns of deep learning with texts more deeply than those who used the comparison 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，再帰的深層学習であるRNN（Recurrent Neural Network）やLSTM（Long Short-Term Memory）を使用し，テキスト集合

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

の学習によって構築されたネットワークを HMM(Hidden Markov Model) に当てはめ，ネットワークの重みの値から，入力層に時系列順に入力される特徴量（本研究ではテキストを構成する単語）の尤度を算出する．その上で，単語の時系列パターンを尤度順に取り出すことで，再帰的深層学習の学習済みネットワークに蓄積された情報を，分類パターン（単純な単語の順序列）として抽出し、その解釈を支援する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 章で HMM を利用した深層学習による分類パターンの抽出・可視化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．

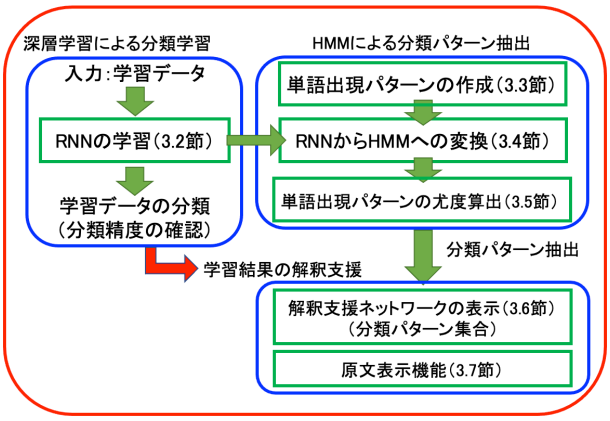

図 1: システムの構成

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，深層学習には判断根拠のブラックボックス問題が存在している．深層学習は非常に複雑なプロセスによって情報を学習し，高い精度で予測・分類を行える．しかし，そのプロセスの複雑さにより，人間が深層学習の判断基準を説明することが非常に難しい．

深層学習のモデルへの信頼性・公平性の説明や判断基準への理解を重視した研究分野として，XAI（Explainable AI：説明可能な AI）[4] が注目されてきている．XAI の研究としては，深層学習モデルの動作を理解・信頼するために，何を学習したか説明を行うことの必要性の提唱 [5, 6] から始まり，実際に，モデル内のデータや変数間の相関関係から動作の説明を試みたり [7]，反事実的条件文を用いてモデルの動作をユーザに理解させる研究 [8] 等，モデルの動作自体を説明できないか試みる研究が行われている．また，モデルの動作の解釈だけではなく，モデルの動作の安定性・信頼性に注目し，悪意のあるデータへの対策 [9] や，モデルの動作を別の論理回路や決定木に当てはめ，モデルの動作やその安定性を評価する研究 [10, 11] も存在する．

そこで，テキストベースの DNN（Deep Neural Network）について，層ごとの学習の流れを単語情報でラベル付けして可視化し，分類基準を人間が理解できる学習ネットワークの解釈支援システムの研究 [12] が存在し，一定の成果が確認された．一方で，上記の研究では深層学習モデルが DNN だったため，単語の時系列情報が失われ，その解釈も一定までしか得られない問題があった．

本研究ではこのような問題意識の下，文章の分類問題を例とし，深層学習モデルとして RNN を使用することで，単語の時系列情報を含めた分類パターンからの分類基準の解釈支援システムの開発を目指す．

3 HMM を利用した再帰的深層学習ネットワークからの分類パターン抽出・可視化システム

本章では，テキストベースの分類タスクの深層学習ネットワークにおいて，HMM を利用した分類パターンの解釈支援を目的としたシステムの構成とその詳細について述べる．

3.1 提案システムの構成

提案システムでは，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を HMM に当てはめ，提案システムの分類パターンの抽出処理部によって各出力 (分類先) を導く分類パターンの尤度に基づく抽出を行う．最後に，システムの利用者は，システムの可視化処理部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．分類パターンを可視化する．また，システムでは分類パターンの意味を理解しやすくするための機能（解釈支援機能）を利用できる．

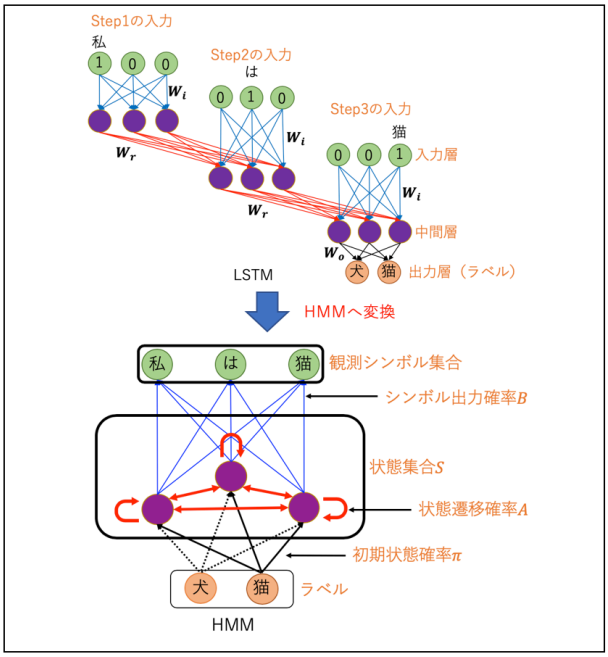

図 2: RNN から HMM への変換

3.2 深層学習による学習ネットワークの形成

3.2.1 テキスト中の単語のベクトル化

深層学習で学習を行う前に，テキストデータはテキスト中の単語を取り出したあと，単語を One hot 法 [13] と呼ばれる手法に従い単語ベクトルの羅列に直す．そして，テキスト中の各単語をその単語ベクトルに置き換え，深層学習への入力データとする．

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，分類先ごとにラベル付けされたテキストデータは，RNN でそれぞれの出力層ノード（分類先）を導くネットワークへの重み付けがされていく．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，LSTM での分類時は，最後の単語が入力されたタイミングで，出力層から分類結果が出力される．

3.3 HMM を用いた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

3.3.1 RNN から HMM への変換

提案システムの分類パターンの抽出処理では，RNN によって得られた学習ネットワークを図 2 のように，ひとつの HMM として処理を行う．

まず，分類パターンの候補として，RNN への入力に使用した全単語の組み合わせを作成する．この時，組み合わせの条件として以下を満たす単語列を候補とする．

- 分類パターン候補の長さ（単語数）は任意で決めた長さで揃えるとする
- 分類パターン候補の単語の順序は実際のテキスト中の単語の出現順序に基づくものとする

次に RNN の入力層ノードを HMM の観測シンボル集合，中間層ノード（LSTM ユニット）を状態集合 $S = \{s\}$ とし，同様に中間層の（再帰的処理による）時系列間の重みを状態遷移確率 A，入力層中間層間の重みをシンボル出力確率 B とする．そして，中間層出力層間の重みを初期状態確率 π とするが，この π はその時選択するラベル（分類先）によって変わる．この時，観測シンボルによる観測系列（前述した分類パターン候補）を $O = o_1, o_2, ..., o_T$（T は観測系列の長さ（前述した分類パターン候補の長さ））, 状態数（中間層ノード数）を N（状態番号は i, j）と置くと，状態遷移確率 A は式 (1)，シンボル出力確率 B は式 (2)，初期状態確率 π は式 (3) となる．

$$\boldsymbol{A} = \{a_{ij} | a_{ij} = P(s_{t+1} = j | s_t = i)\} (1 \leq i, j \leq N) \quad (1)$$

$$\boldsymbol{B} = \{b_{ij}(o_t) | b_{ij}(o_t) = P(o_t | s_{t-1} = i, s_t = j)\} \quad (1 \leq i, j \leq N, 1 \leq t \leq T) \quad (2)$$

$$\boldsymbol{\pi} = \{\pi_i | \pi_i = P(s_0 = i)\} (1 \leq i \leq N) \quad (3)$$

この時，ある分類先 x に対して，単語出現パターン $\boldsymbol{O}$ が存在する時，初期状態確率を $\boldsymbol{\pi}_x$ と表すと，尤度 $P(O|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B})$ は，式 (4) で算出される．

$$\begin{aligned} P(\boldsymbol{O}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) &= \sum_{all \boldsymbol{S}} P(\boldsymbol{S}|\boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) P(\boldsymbol{O}|\boldsymbol{S}, \boldsymbol{\pi}_x, \boldsymbol{A}, \boldsymbol{B}) \\ &= \sum_{all s_0 \ldots s_T} \pi_{x s_0} a_{s_0 s_1} b_{s_0 s_1}(o_1) \cdot a_{s_1 s_2} b_{s_1 s_2}(o_2) \cdot \\ &\quad \ldots \cdot a_{s_{T-1} s_T} b_{s_{T-1} s_T}(o_T) \end{aligned} \quad (4)$$

最後に，全ての単語出現パターンについて式 (4) で尤度を算出し，尤度の高い順に，単語出現パターンを分類に寄与する分類パターンとして抽出する．

3.3.2 分類先を導く分類パターンの可視化

提案システムの可視化処理部では，分類先に強く結びつく，尤度の高い分類パターン集合が表示される．例として，5 種類の和菓子の作り方に関するテキスト集

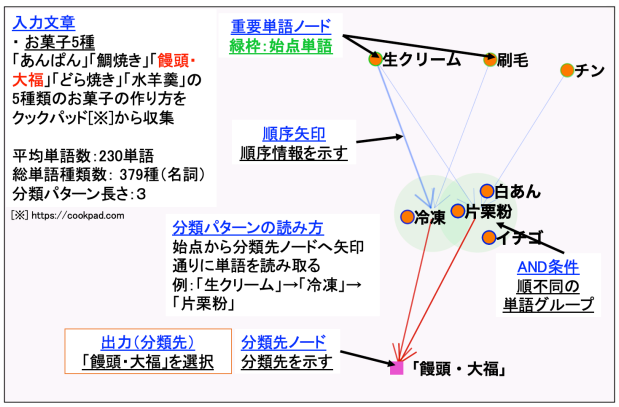

図 3: 提案システムの画面例

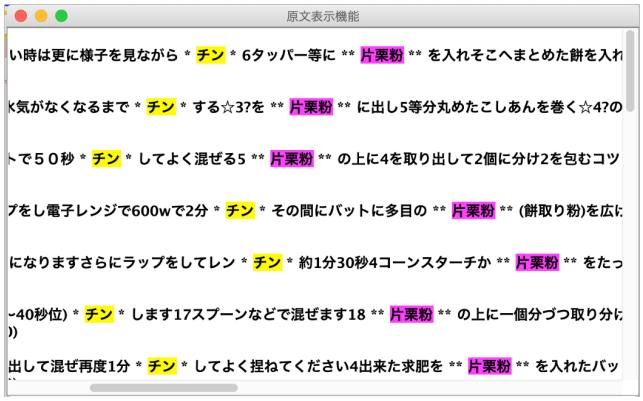

図 4: 原文表示例(「饅頭・大福」についてのテキストに対して単語「チン」と「片栗粉」を選択)

合[1]の分類を行った場合の，提案システムのメイン画面を図 3 に示す．表示分類先は「饅頭・大福」を選択している．図 3 では，分類パターン中の単語の流れを矢印の向きで表し，分類パターンを構成する単語をノードで表している．また，尤度の大きさを矢印の太さで表している．

3.4 分類パターン解釈支援機能：原文表示機能

システムには，利用者が抽出された分類パターンの解釈行いやすいように，その表示内容の補足を行う機能がある．その中のひとつである原文表示機能について述べる。

分類パターンの解釈に向けて，単語情報だけでは，その単語が実際にどのような文脈で使われていたのかを把握することは難しい．そのため，原文表示機能により，分類パターン中の単語群が，実際に学習に用いたテキスト内でどのように使われているかを表示する．

[1] クックパッド (URL:https://cookpad.com) から収集

ユーザは，解釈支援ネットワーク上で単語を選択することで，原文の中でその単語を含む文章が表示され，参照することができる．ただし，見やすさを考慮して，表示されるのは 1 つの文章につき，選択した単語の前 10 単語，後 10 単語までの区間とした．また，単語は最大 2 種類まで選択でき，その場合は単語間の文章は全て表示される．図 4 に 5 種類の和菓子の作り方について分類先「饅頭・大福」のテキストを使用し，単語「チン」と「片栗粉」を順番に選択した時の分類パターンの原文表示例を示す．

4 HMM を適用したテキスト分類パターン解釈支援システムの有効性の検証実験

本章では，深層学習の深い知見を有さない被験者が提案システムの出力する単語の出現パターンをもとに，分類パターンの解釈を行うことができるかを検証した実験について述べる．

4.1 実験手順

実験は，課題 1 「キャラセリフ」、課題 2 「家電レビュー」、課題 3 「ゲームレビュー」の 3 つの課題（詳細は下記参照）に対して，各課題ごとに指定する「出力ラベル」に分類される文章の分類パターンの解釈を行ってもらった．実験は深層学習についての深い見識がない 16 名の大学生，大学院生に対して行い，提案システムを用いるグループと，比較システムを用いるグループに 8 名ずつの 2 つに分けて行った．

実験では，学習結果を言葉で表して説明する提案システムとの比較として，分類先カテゴリに特有の単語を抽出する TFIDF から，分類先の特徴を解釈するシステムを用意した．TFIDF 値を利用する比較システムを採用したのは，TFIDF による特徴的な単語単体や，単語の組合せから解釈した結果と，提案システムによる単語の時系列関係を参照した解釈を比較することで，分類パターンの解釈における単語の時系列関係を提示することの有効性を明示できると考えたためである．

提案システムを用いるグループでは，提案システムを用いて，分類に寄与する単語（単語単体，組合せ，時系列順序）を見つけてもらった．比較システムを用いるグループでは，比較システムとして，指定する出力ラベルに特有の単語を式 (5) の TFIDF 値により抽出してリスト形式で提示するシステムを用意し，これらの単語を元に，分類に寄与する単語を見つけてもらった．また比較システムにおいても，提案システムの原文表示機能を利用できるようにした．

ある単語 i の $TFIDF_i$ = 単語 i の文章頻度

$$\times(\log(\frac{\text{出力ラベル数}}{\text{単語 } i \text{ の } DF \text{ 値}})+1) \quad (5)$$

実験手順について，以下のステップで両グループの被験者に解釈を行ってもらった．その際，提案システムで表示される分類パターン数は，単語数 3 個から構成される分類パターンを尤度の高い順に 5 つとした．また，比較システムでの表示単語数も提案システムに合わせて 15 個とした．

手順 1 解釈対象の出力ラベルを選択する：課題 1 では「ツンデレ」、課題 2 と課題 3 では「役に立つ」に分類されるレビューを対象とした．

手順 2 それぞれの選択した出力ラベルに対応した「解釈の目的」を読んで内容を理解する．

手順 3 選択した出力について，「解釈支援ネットワーク」を表示させて，出力に寄与すると思われる特徴（単語単体や組合せ，時系列順序等）を 10 個見つける．

手順 4 注目した特徴に対して，原文表示機能を用いながら考案してもらう．

学習データの詳細について述べる。課題「キャラセリフ」においては，Twitter の「ツンデレ bot」「デレデレ bot」「キャラセリフ bot」からそれぞれ「ツンデレ」「デレデレ」「ノーマル」のキャラの特徴を持つセリフを 500 件ずつ，計 1500 件を利用し、「ツンデレ」を解釈対象とした．課題「家電レビュー」と課題「ゲームレビュー」においては、amazon の「人気家電製品」の上位 50 種類と「人気ゲームソフト」の上位 100 種類より、「役に立つ」（星 4 以上かつ,「役に立つ人数」が 10 以上）レビュー,「役に立たない」（星 4 以上かつ,「役に立つ人数」が 0）レビュー,「低評価」（星 2 以下）レビューをそれぞれ課題「ゲームレビュー」では 1036 件ずつ，計 3108 件を利用し、課題「ゲームレビュー」ではそれぞれ 1473 件ずつ，計 4419 件を利用した。また、解釈対象は「役に立つ」を選択した．

学習は LSTM によって行い，中間層は 1 層とした．中間層のノード数は，分類精度が 95%を下回らない範囲で，ノード数を減らす操作を行った．学習率は 0.1，l1 ノルム係数，l2 ノルム係数はともに 0.0001，学習回数は 50 回で学習を行った．

4.2 結果と考察

まず，被験者により記述された解釈の妥当性の内訳（被験者平均）を図 5 に示す．ただし，解釈の妥当性の内訳は，以下に定義する内容をもとに，著者の 1 名が分類を行った．

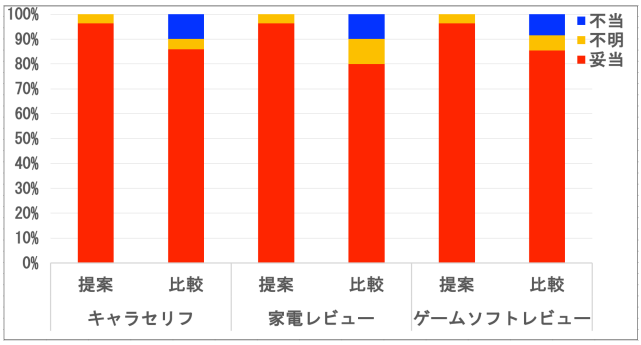

図 5: 被験者の解釈の妥当性の内訳（被験者平均）

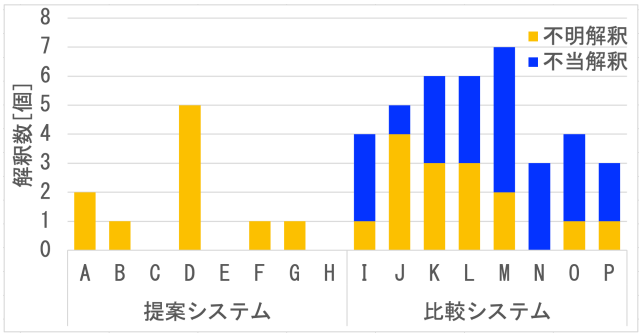

図 6: 被験者別の不明，不当解釈個数

- 妥当な解釈（妥当）：内容の正しさが原文から確認でき,「解釈の目的」にも合っている．
- 妥当か判断できない解釈（不明）：内容の意図がはっきりせず，妥当か妥当でないかが判断できない．
- 妥当でない解釈（不当）：解釈の内容に誤りが確認できたか．または「解釈の目的」に合った内容ではない．

図 5 の結果から，提案システムでは，97%以上の解釈が妥当な解釈に分類され，その正しさが確認できた．特に，比較システムで 10%近く存在している妥当でない解釈については，提案システムの結果ではひとつも見られなかった．また，妥当か判断できない解釈についても，比較システムでは全体の 5%から 10%程度に含まれていたが，提案システムでは全体の 3%以下であった．このことから，提案システムではより意図が明確で妥当な内容の解釈が行われていたと言える..

また，被験者ごとの「妥当でない解釈（不当解釈）」と「妥当かどうか判断できない解釈（不明解釈）」の数を図 6 に示す．図 6 の A から H は提案システムの被験者 8 名，I から P は比較システムの被験者 8 名を表す．

図 6 より,「不明解釈」を行った被験者の数は，提案システムで 5 人，比較システムで 7 人と大きな差はなかった一方で,「不当解釈」を行った被験者は提案シス

テムで0人に対し，比較システムで8人全員となり，1人を除いて複数の「不当解釈」を与えていたことがわかる．そのため，個人差によらず提案システムを用いた方が，より妥当な解釈を与えられたことが確認できる．

以上をまとめると，提案システムでは，比較システムより，より正解率の高い，典型的で妥当な解釈が導き出せることが確認できた．これは，特に複数の単語の時系列関係に注目して解釈が行えることが要因と言える．

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時系列関係が学習できる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解釈支援システムを提案した．本研究の特徴として，学習済みの再帰的深層学習のネットワーク構造をHMMに当てはめて処理することで，モデルの構造を変えることなく，容易に，学習された特徴量の時系列情報を抽出できる点が挙げられる．提案システムの有効性を確かめる検証実験では，提案する環境が，深層学習に精通していないユーザでも，時系列情報を含む分類パターンから，容易に原文の内容を広くカバーする妥当な解釈につなげられることを確認した．

今後は，BERT（Bidirectional Encoder Representations from Transformers）などのさらに複雑な深層学習ネットワークを対象とした解釈環境を構築することを目指す．

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] D. Gunning, ‘Explainable artificial intelligence (xAI)’, Tech. rep., Defense Advanced Research Projects Agency (DARPA), 2017.

[5] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, ‘Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?’, IEEE Computational Intelligence Magazine 14 (1), 69-81, 2019.

[6] J. Haspiel, N. Du, J. Meyerson, L. P. Robert Jr, D. Tilbury, X. J. Yang, A. K. Pradhan, ‘Explanations and expectations: Trust building in automated vehicles’, Companion of the ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp. 119-120, 2018.

[7] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, M. Sebag, ‘Learning functional causal models with generative neural networks’, Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, pp. 39-80, 2018.

[8] R. M. J. Byrne, ‘Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning’, Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pp. 6276-6282, 2019.

[9] X. Yuan, P. He, Q. Zhu, X. Li, ‘Adversarial examples: Attacks and defenses for deep learning’, IEEE Transactions on Neural Networks and Learning Systems 30 (9), 2805-2824, 2019.

[10] G. Audemard, F. Koriche, P. Marquis, ‘On Tractable XAI Queries based on Compiled Representations’, KR Proceedings 2020 Special Session on KR and Machine Learning, pp.838-849, 2020.

[11] Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, ‘Interpreting CNNs via decision trees’, IEEE Conference on Computer Vision and Pattern Recognition, pp. 6261-6270, 2019.

[12] 安藤雅行，河原吉伸，砂山渡，畑中裕司，‘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’，知能と情報 (日本知能情報ファジィ学会誌)，Vol.31, No.4, pp.779-787, 2019.

[13] 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