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23回)
SIG-AM-23-08

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

Extraction and Visualization of Classification Patterns from Deep Learning Networks using HMM

安藤 雅行[1,2*]　河原 吉伸[2,3]　砂山 渡[4]　畑中 裕司[4]
Masayuki ANDO[1,2]　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] 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

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 extracts classification patterns from the trained learning networks of LSTM using HMM. 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 the weights of the classification patterns extracted from data given a unique pattern. The results showed that the proposed system can extract classification patterns effective for interpretations of the learning networks.

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（主に LSTM(Long short-term memory)）を使用し，テキスト集合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，ネットワークの層に付けられた重みの値から，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．そして，その単語の順序を考慮した組み合わせ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

*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

れた情報を，分類パターン（単純な単語の順序列）として抽出することができ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．

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，構造は単純だが特徴量（テキスト中の単語）の有無だけ学習し，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るようにした．また，深層学習が持つ，ネットワークの各層ノード間の関係にマルコフ性が存在する [11] 性質を利用して，学習ネットワークを HMM に当てはめることで，ネットワークの各層に付けられた重みの値から，タイムステップごとの入力単語の順序列に対する尤度を算出，そこから，尤度の高い単語の順序列を分類のパターン，つまり出力を導くルールとして抽出するシステムの構築を目指す．また，システムでは抽出された分類パターンを可視化するインターフェースを備えている．

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，本研究で開発した 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抽出・可視化システム（以後，提案システム）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 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 LSTM 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．

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 [12] 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．

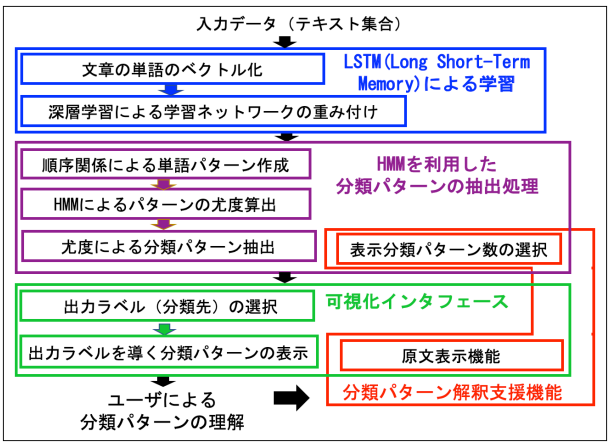

図 1: 提案システムの構成

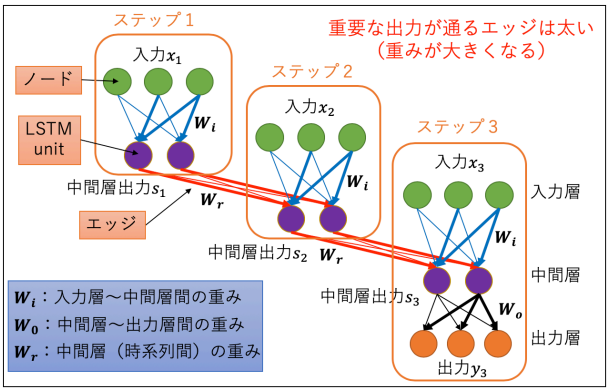

図 2: 再帰的深層学習の学習ネットワークと学習の様子

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，LSTM 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．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，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 LSTM の 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，LSTM によって得られた学習ネットワークを図 3 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，LSTM 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

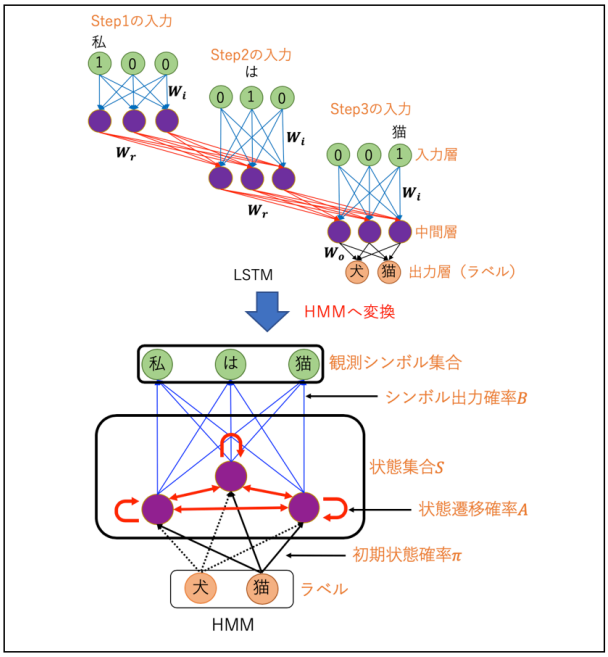

図 3: LSTM の HMM への変換

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 LSTM 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．

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

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

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

この時，A，B，π で構成される HMM を式 (4) のように略記する．

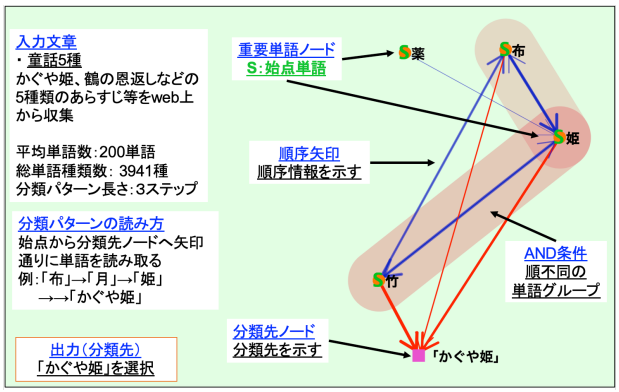

図 4: 提案システムの画面例

$$\lambda = \{\pi, A, B\} \tag{4}$$

最後に，分類パターン候補の尤度を算出し，尤度の高い順に分類パターンとして抽出する．あるラベル a に対して，分類パターン候補 $O = o_1, o_2, ..., o_T$ がある時，分類パターン候補 O の尤度 $P(O|\lambda)$ は，式 (5) で算出される．

$$\begin{aligned} P(O|\lambda) &= \sum_{allS} P(O|\lambda)P(O|S,\lambda) \\ &= \sum_{all s_0 \ldots s_T} \pi_{s_0} a_{s_0 s_1} b_{s_0 s_1}(o_1) \cdot a_{s_1 s_2} b_{s_1 s_2}(o_2) \cdot \\ &\quad \ldots \cdot a_{s_{T-1} s_T} b_{s_{T-1} s_T}(o_T) \end{aligned} \tag{5}$$

こうして LSTM の学習ネットワークから，時系列を考慮した単語の並びとしての分類パターンを抽出し，各分類先への寄与の強さを尤度で表すことができる．なお，式 (5) の計算量は $O(2TN^T)$ であるため，深層学習モデルが大きくなると計算量が非常に大きくなってしまう．そのため，本来は forward-backward アルゴリズム等を用いるが，今回使用したモデルは比較的規模が小さいため，式 (5) で計算を行った．

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 種類の童話のあらすじに関するテキスト集合の分類を行った場合の，提案システムのメイン画面を図 4 に示す．表示分類先は「かぐや姫」を選択している．図 4 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．

表 1: 学習ネットワーク表示機能

機能名	効果
表示分類パターン数の増減	一つの分類先につき，いくつの分類パターンを表示するか選択する
分類パターン長の選択	分類パターン候補作成時に，分類パターン候補の長さを選択する
原文表示機能	分類パターン（順序を考慮した単語組み合わせ）が，原文中でどのように出現しているかを表示する

表 2: テキストデータの詳細

データ名	内容
人工データ（6 パターン）	3 種類の重要記号「A0」「A1」「A2」を 1 つずつ使った組み合わせ 6 パターン（例「A0」→「A1」→「A2」,「A1」→「A2」→「A0」など）をラベルとし，重要記号間にランダムで選択したノイズ記号「N0」から「N999」を 5 つづつ挿入したものを 1 テキストとして，パターンごとに 5,000 (計 30,000) テキスト用意した
人工データ（24 パターン）	4 種類の重要記号「A0」「A1」「A2」「A3」を 1 つずつ使った組み合わせ 24 パターン（例「A0」→「A1」→「A2」→「A3」,「A1」→「A2」→「A3」→「A0」など）をラベルとし，重要記号間にランダムで選択したノイズ記号「N0」から「N999」を 5 つづつ挿入したものを 1 テキストとして，パターンごとに 1,000（計 24,000）テキスト用意した

3.4 分類パターン解釈支援機能

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，その表示内容を調整できる機能がある．その主なものを表 1 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，提案システムについて，分類パターンの抽出に焦点を当て，抽出された分類パターンが，テキストデータの理解（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「順序を考慮した任意の長さの異なる単語の組み合わせ」と定義する．

表 3: 学習 LSTM 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 LSTM ユニット数	出力層ノード数
人工データ（6 パターン）	1,003 記号	1,003	10	6
人工データ（24 パターン）	1,004 記号	1,004	10	6

表 4: 人工データ（6 パターン）の偏差値

ラベル	尤度 10^{-3}	平均 10^{-3}	偏差値	標準偏差 10^{-3}
A0 → A1 → A2	4.56	0.10	149.4	0.45
A0 → A2 → A1	5.20	0.10	149.5	0.51
A1 → A0 → A2	4.75	0.10	149.4	0.47
A1 → A2 → A0	5.39	0.12	149.4	0.53
A2 → A0 → A1	5.06	0.11	149.4	0.50
A2 → A1 → A0	4.70	0.10	149.4	0.46
平均	4.94	0.11	149.4	0.49

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して，決まったパターンを持つ記号の組み合わせで構成された人工データを 2 つ用いる．データの詳細について表 2 に示す．

続いて，深層学習として使用した LSTM モデルの概要を表 3 に示す．また，抽出する分類パターンは長さ 3 で個数はラベル（分類先）ごとに 100 とする．なお，２つの人工データに対する LSTM の分類精度はどちらとも 100%である．

4.2 実験手順

実験は著者 1 名で行い．テキスト「人工データ（6 パターン）」,「人工データ（24 パターン）」について，提案システムの分類パターン抽出処理部によって，各分類先ごとに尤度の高い順に 100 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の尤度の偏差値や標準偏差を算出し，決められたパターンを持つ分類パターン（各ラベルごとに決めたパターンのうちノイズを含まないもの）とその他のノイズを含む分類パターンの偏差値等を比較した．

なお，分類パターンの長さは 3 なので，テキスト「人工データ（6 パターン）」では決められたパターンを持つ分類パターンは各パターンに 1 つずつ（決められたパターンの長さが 3 のため）となり，テキスト「人工データ（24 パターン）」では，決められたパターンを持つ分類パターンは各パターンに 4 つずつ（決められたパターンの長さが 4 のため）となる．

4.3 結果と考察

テキスト「人工データ（6 パターン）」について抽出された分類パターンの尤度上位 100 の偏差値を出し，そのうち，各ラベルごとに決められたパターンを持つ分類パターンの偏差値，100 個中の尤度の平均，標準偏差を表 4 に示す．また，テキスト「人工データ（24 パターン）」について抽出された分類パターンの尤度上位 100 の偏差値を出し，そのうち，各ラベルごとに決められたパターンを持つ分類パターンの偏差値，100 個中の尤度の平均，標準偏差を表 5 に示す．ただし,「人工データ（24 パターン）」の方は決められたパターンを表す分類パターンが各ラベルごとに 4 つずつあるため，偏差値はその 4 つの平均とする．

表 4，表 5 より，決められたパターンを持つ分類パターンは，その他のノイズを含む分類パターンより尤度や偏差値が大きくなることがわかった．これは，提案システムでは，テキスト特有の特徴を表す，学習ネットワークの解釈に適した分類パターン（人工データにおける決められたパターンを持つ分類パターン）が，その他の学習ネットワークの解釈に適さない分類パターン（ノイズ等を含む，不必要な情報を持った分類パターン）より明確な差を持って抽出できることを示す．これにより，抽出する分類パターンの尤度に閾値を設けて可視化すれば，提案システムの利用者は，テキストの分類に明確に寄与する分類パターンのみを対象とした解釈ができる．

また，表 4 と表 5 を比較すると，表 4 では，決められたパターンを持つ分類パターンの尤度が平均の 10 倍前後になっており，偏差値も 150 近い値となっている一方，表 5 では，決められたパターンを持つ分類パターン尤度が平均の 2 倍から 3 倍程で，偏差値も 100 を下回っている．これは，表 4 の決められたパターンを持つ分類パターンは，完全にその分類先にしか存在していないので尤度が高くなるが，表 5 の決められたパターンを持つ分類パターン 4 つは，一つひとつは他の分類先にも存在しているので，表 4 より尤度が高くならなかったからと思われる．よって，学習させるテキストの長さに適した長さの分類パターンを選ぶ方が，より尤度の高い特徴的な分類パターンを抽出できると考えられる．

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 LSTM で行い，

表 5: 人工データ（24 パターン）の偏差値

ラベル	尤 度 10^{-7}	平 均 10^{-7}	偏差値	標準偏差 10^{-7}
A0 → A1 → A2 → A3	4.76	1.66	93.1	0.72
A0 → A1 → A3 → A2	4.93	1.84	93.8	0.71
A0 → A2 → A1 → A3	5.36	1.96	91.2	0.82
A0 → A2 → A3 → A1	4.71	1.90	88.0	0.74
A0 → A3 → A1 → A2	5.28	1.92	93.7	0.77
A0 → A3 → A2 → A1	4.32	1.66	90.6	0.66
A1 → A0 → A2 → A3	5.41	1.78	93.4	0.84
A1 → A0 → A3 → A2	5.21	1.77	94.0	0.78
A1 → A2 → A0 → A3	5.32	1.74	93.5	0.82
A1 → A2 → A3 → A0	5.65	2.02	92.7	0.85
A1 → A3 → A0 → A2	4.73	1.67	93.5	0.70
A1 → A3 → A2 → A0	4.45	1.54	93.3	0.67
A2 → A0 → A1 → A3	5.45	1.92	92.2	0.84
A2 → A0 → A3 → A1	4.70	1.79	91.0	0.71
A2 → A1 → A0 → A3	5.58	2.06	92.7	0.82
A2 → A1 → A3 → A0	4.55	1.79	92.5	0.65
A2 → A3 → A0 → A1	5.53	1.80	93.7	0.85
A2 → A3 → A1 → A0	5.55	2.12	92.1	0.81
A3 → A0 → A1 → A2	5.32	1.79	94.3	0.80
A3 → A0 → A2 → A1	3.83	1.44	91.2	0.58
A3 → A1 → A0 → A2	5.93	2.10	93.6	0.88
A3 → A1 → A2 → A0	5.24	1.83	92.5	0.80
A3 → A2 → A0 → A1	5.54	1.75	93.4	0.87
A3 → A2 → A1 → A0	4.92	1.72	94.4	0.72
平均	5.13	1.77	93.2	0.77

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，複雑な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，容易に，学習された特徴量の順序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，提案システムで抽出された分類パターンの尤度について，決められたパターンを表す分類パターンとノイズを含む解釈に向かない分類パターンを比較し，決められたパターンを表す分類パターンの方が，尤度が特徴的に大きく抽出できていることから，提案システムで抽出される分類パターンはテキストの特有の特徴を有しているものが表示されやすいと結論づけた．今後の研究では，実際に抽出された特徴を用いて，それらからどのような学習ネットワークの解釈が行えるかを重視して進めていく予定である．

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，“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

[11] R. Shwartz-Ziv and N. Tishby. “Opening the black box of deep neural networks via information”, arXiv preprint arXiv:1703.00810, 2017

[12] 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