CNN を用いたテキスト分類モデルの可視化と単語の影響度分析

森　のどか　　指導教員：小林　邦和

1　はじめに

近年人工知能の技術が年々向上し日常でもよく見聞きするようになった．人工知能の 1 つである深層学習モデルは現在盛んに研究が行われており，多くの分野で高い精度を発揮している．その反面，ブラックボックス問題というモデルが正しい結果を推論したとしても，なぜそのような結果となったのか根拠を説明できないという問題がある．つまり医療の現場や刑事司法等，人の命や人生に関わる場面などでは得られた結果に対し根拠がないため鵜呑みにできず，安易な深層学習モデルの導入に対し懸念の声が上がっている．そのため深層学習モデルに推論結果の根拠を示す解釈性付与の必要性が高まっている [1].

深層学習モデルに解釈性を付与する研究は特に画像認識タスクにおいて多く行われており，テキスト分類タスクに解釈性を付与する研究はあまり多く行われていない．対話システムの制御やニュース記事のジャンル振り分け等にテキスト分類タスクの活用が期待されているため，テキスト分類タスクにおいても解釈性を付与させることが必要であると考える．

このような背景から本研究ではテキスト分類タスクにおいてモデルの分類結果に対する根拠の可視化を目指す．具体的には，畳み込みニューラルネットワークを用いてテキスト分類を行い，テキスト中の単語が分類結果に与える影響度を重みパラメータの微分値で定量化し根拠を可視化する．また求めた影響度と実際の単語の持つ意味との相違を検証するため，単語の感情辞書と比較分析する．

2　提案手法

本研究ではまずテキスト分類を行うモデルとして TextCNN を利用する．TextCNN に対しテキスト中の単語が分類結果に与える影響度を算出し，根拠の可視化を行う．そして求めた影響度が実際の単語の意味とどれくらい異なるのかを確かめるため，単語の感情度と相関係数を用いて確認する．提案手法の流れを図 1 に示す．

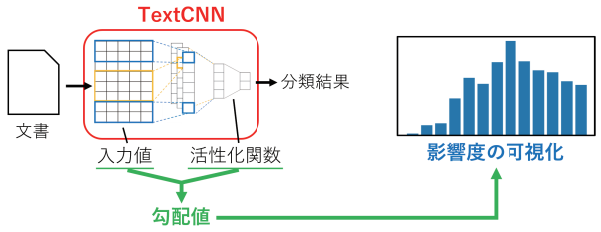

図 1　提案手法の流れ

2.1　TextCNN

TextCNN は画像認識タスクで高い精度を誇っている畳み込みニューラルネットワークをテキスト分類タスクに転用したモデルである [2]．構造は図 2 のようになっている．

TextCNN はまず入力された文書に対して word2vec という埋め込み処理を行うことで数値化をする．ここでデータ中の文書の長さはそれぞれ異なるため，対象の文書の中で最も単語数の多い文書の単語数に合わせて『unknown』を意味する『⟨UNK⟩』でパディングを行う．

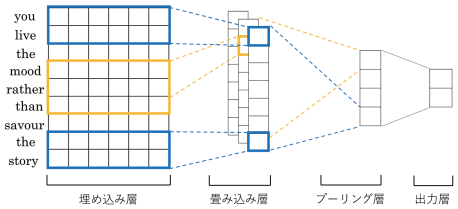

図 2　TextCNN の構造

次に畳み込み層ではフィルタをかけることで単語の特徴マップを求める．TextCNN では複数の特徴を取得するためにウインドウサイズの異なる複数のフィルタを用いる．次に各フィルタをかけて生成された特徴量にプーリング演算を適用し，特徴として重要な情報を残しながら圧縮する．そして全結合層を経て，softmax 関数をかけ各クラスに属する確率を出力する．

2.2　微分値を利用した可視化

この手法は主に CNN の画像認識タスクにおいて用いられている [3]．CNN を用いた画像認識を行う場合，各クラス c の活性化関数 S_c を計算し，最終的に出力される分類結果 $class(x)$ は式 (1) より求められる．

$$class(x) = \arg\max_{c \in C} S_c(x) \tag{1}$$

関数 S_c が部分的に微分可能な場合，任意の入力値 x に対して微分値を求めることで，入力画像の各画素が分類結果へ与える影響度を求めることができる．微分値が大きいほど，分類結果に与えた影響度が大きいことを意味している．この微分値を用いて感度マップを作成することで視覚的にそのモデルが分類した根拠を確認することができる．図 3 は画像認識モデルにドーベルマンの画像を入力し，正しく分類したときに微分値が低い画素を黒くしてマッピングした図である．

図 3　画像認識モデルに対して微分値を用いてマッピングした例

本研究ではこの手法をテキスト分類における根拠の可視化に用いた．

3　計算機シミュレーション

3.1　目的

本研究では以下の 2 点の目的に沿ってシミュレーションを行う．

1. モデルの分類結果に対する根拠の可視化
2. 単語の影響度分析

まず TextCNN に対し入力された単語に対する微分値を用いて根拠の可視化を行う．そして単語ごとの影響度と感情度の相

関係数を求めることで相違を定量化する．感情度は感情辞書に記載されている単語ごとにその単語が表す感情の度合いを数値化したものである．

3.2 使用データセット

3.2.1 学習用データセット

今回モデルの学習データとして Bo Pang らが作成した Movie Review Data を使用した [4]．このデータセットは映画のレビューがまとめられており，5331 件のポジティブなレビューと 5331 件のネガティブなレビューの構成となっている．

3.2.2 分析用データセット

感情度は SentiWordNet という感情辞書を用いた [5]．この感情辞書は大量の英単語にポジティブ，ネガティブ，客観性のスコアが合計 1 になるように割り振られているデータセットである．SentiWordNet の一部を表 1 に示した．今回はこの内ポジティブスコア [PosScore] とネガティブスコア [NegScore] を用いた．

表 1　SentiWordNet の一部

POS	PosScore	NegScore	SynsetTerms
a	0.125	0	able#1
a	0	0.75	unable#1
n	0	0	entity#1
v	0.125	0	respire#2

3.3 シミュレーション環境

TextCNN に対し Movie Review Data でモデルを学習し，入力された文書がポジティブかネガティブか分類するモデルを作成する．本研究でモデルを学習させたときのパラメータは表 2 の通りである．

表 2　学習時のパラメータ設定

parameter	value
learning rate	0.0001
Regularization(Lasso) coefficient	0.05
dropout rate	0.5
window size	3, 4, 5

3.4 分類結果に対する根拠の可視化

学習終了時のモデルの検証データに対する accuracy の値は 0.74，テストデータに対する accuracy は 0.75 となった．次に学習させたモデルに対し，テストデータを入力したときの単語が分類結果に与えた影響度を計算する．ポジティブな文書を入力し正しく分類できた際の単語ごとの微分値と，ネガティブな文書を入力し，正しく分類できた際の単語ごとの微分値の上位それぞれ 5 単語を示したのが表 3，4 である．この表において，全テストデータの中で出現回数が 3 回以上の単語のみをまとめている．

表 3　ポジティブな文書を入力し正しく分類したときの単語の微分値

単語	微分値の平均
affecting	6.069
provides	6.015
amusing	5.969
captivating	5.875
solid	5.861

表 4　ネガティブな文書を入力し正しく分類したときの単語の微分値

単語	微分値の平均
incoherent	6.071
baked	6.063
thrown	5.866
conceived	5.851
generic	5.782

3.5 単語の影響度分析

次に前節で求めた影響度と，単語が元々持つ感情度との比較を行った結果を示す．影響度と，感情度の相違を確かめるために相関係数を求めた．

ポジティブの影響度とポジティブの感情度の相関係数は 0.0816 となり，ネガティブの影響度とネガティブの感情度の相関係数は 0.0220 となった．

3.6 考察

根拠の可視化を行った結果，微分値が高い単語には実際にポジティブな意味を持つ単語，ネガティブな意味を持つ単語が多く，感情度を学習できている単語もあることが分かった．その一方微分値が高い前置詞もあり，前後の単語のつながりも関係していると考えられる．

また 2 つの相関分析の結果から，どちらの相関係数も絶対値が 0.2 を下回ったためほとんど相関がないことがわかった．このことから単語全体で見ると，単語の影響度と感情度の大小は関連していないことが分かった．

4 おわりに

本研究では CNN を用いたテキスト分類モデルに対し，入力された単語に対する分類結果の微分値を求めることで影響度を可視化した．そして求めた影響度と実際の単語の感情度との相違を確かめるため，感情辞書の感情度との相関を分析を行った．また，入力する単語が単語のつながりによって影響度が変化するか確かめた．

先行研究では CNN を用いた画像認識における根拠の可視化に対してノイズを減らす工夫がされた研究や，ニューラルネットワークの中間層を可視化する研究などがある．今後の課題としてそれらの手法の導入を目指していきたい．

参考文献

[1] 瀬光孝之, 吉村玄太, 毬山利貞, 杉本和夫: 「解説 機械学習モデルの解釈性に関する最新動向」電子情報通信学会誌, Vol.102, No.10 (2019)

[2] Yoon Kim:「Convolutional Neural Networks for Sentence Classification」EMNLP, pp.1746-1751 (2014)

[3] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, Martin Wattenberg 「SmoothGrad: removing noise by adding noise」CoRR, Vol.abs/1706.03825 (2017)

[4] Bo Pang, Lillian Lee: 「Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales」Proceedings of the ACL (2005)

[5] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani「SENTI WORD NET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining」LREC'10 (2010)