深層学習による文書分類への解釈性の付与に関する一考察

中村 鴻介†1　山口 実靖†2

†1, †2 工学院大学大学院 工学研究科 電気・電子工学専攻 〒163-8677 東京都新宿区西新宿 1-24-2

E-mail: †1 cm19037@ns.kogkauin.ac.jp, †2 sane@cc.kogakuin.ac.jp

あらまし 深層学習は従来の機械学習手法と比較すると，自然言語処理などの様々な問題に対し高精度な推論を達成できる手法として期待されている．しかし，深層学習などの機械学習には推論結果についての説明性や推論モデルに対する解釈性がないものが多いという問題点も指摘されている．本稿では，機械学習の推論結果に解釈性を付与する手法について考察をする．まず，SVM または DNN を用いた文書分類モデルに解釈性を付与する既存手法を紹介する．次に，attention 機構を組み込んだ LSTM を用いた文書分類モデルに解釈性を付与する手法を提案し，それらによる文書分類の解釈性付与結果を示す．

キーワード 機械学習，深層学習，深層学習の解釈性，LSTM, Attention

1. はじめに

近年，深層学習が機械学習手法の一つとして注目されている．深層学習はニューロンを多層に組み合わせて構成される深層ニューラルネットワーク(DNN; Deep Neural Network)，畳み込みニューラルネットワーク(CNN; Conventional Neural Network)や再帰ニューラルネットワーク(RNN; Recurrent Neural Network)などが提案されている．これらは従来の機械学習手法に比べて，言語認識や画像認識などに対して高精度な分類や推論が可能であり，様々な分野で応用されている．しかし，文献[1][2]などで，深層学習をはじめとした多くの機械学習アルゴリズムには推論結果に対する解釈性や説明性がないという指摘がされている．しかし，裁判における判決や経営判断など責任のある決定をする状況においては，説明性や解釈性の付与が求められる状況も多く，そのような状況においては推論結果についての説明性や解釈性の付与が重要であると考えられる．

また一方で，オンラインショッピングサイトの普及によって，インターネット上にはある対象についての主観的評価情報が多く存在するようになった．主観的評価対象にはオンラインショッピングサイトの商品などがあり，主観的評価情報には商品レビューなどが挙げられる．このような主観的評価情報を解析し，肯定的なものと否定的なものに分類しこれを提示することは，提供者や利用者双方にとって有益であると考えられる．そして，主観文書を内容に基づいて肯定的または否定的に分類する手法の確立は重要であると考えられる．

我々は過去に文献[3]で，主観文書が肯定的文書であるか否定的文書であるかを分類するモデルを SVM および DNN にて構築し，それらのモデルの重みなどの情報を元に，分類モデルが主観文書のどの語に着目して分類を行っているか明らかにして，解釈性付与の有効性について考察した．しかし，同文献ではシンプルな DNN を用いており, Long Short-Term Memory (LSTM) [4]を用いた検証が行われていない．本稿では，self-Attention[6]による分類モデルを構築し，これらの分類モデルにて主観文書の学習と分類を行う．次に文献[3]にて我々が提案した SVM と DNN での解釈性付与結果を示し，self-attention による分類モデルへの解釈性付与結果を示す. 最後に SVM および DNN と self-attention を比較し考察を行う．

本稿の構成は以下の通りである．2 章で関連研究について述べ，3 章で各分類モデルにおける主観文書分類の学習とその精度を示す．4 章で既存手法による分類モデルへの解釈性付与とその例について述べる．5 章で提案手法による分類モデルへの解釈性付与について述べる．6 章で提案手法による解釈性付与の例を示し，7 章で本稿をまとめる．

2. 関連研究

本章では，深層学習と深層学習の解釈性の関連研究について述べる．2.1 節では，時系列データを扱う際に広く用いられている手法のひとつである LSTM と，近年自然言語処理で広く用いられている手法のひとつである Attention と，それを用いた手法のひとつである self-attention について述べる．2.2 節では深層学習の判断根拠を示す手法について述べる．

2.1. 深層学習

LSTM[4]は RNN では学習が不可能であった長期的な時系列データを，RNN に比べてより良く学習できる学習手法である．LSTM では長期的な時系列データを扱う際に生じてしまう勾配消失や勾配爆発を抑制するために入力ゲート，出力ゲート，忘却ゲートが用いられている．

Attention[5]は LSTM の途中の隠れ状態の値も用いて推論を行うように，翻訳に用いられる Encoder-Decoder

モデルに併用される機構である．LSTM に Attention を併用することによって，Encoder の各シーケンスの中間層をすべて記録し，文脈情報（「今日」が「Today」に対応など）に Decoder を着目させることによって長文翻訳タスクにおける BLEU スコアを向上させた．

図 1 に Attention 付き Encoder-Decoder モデルを示す．図 1 中の$\overline{\boldsymbol{h}_s}$はエンコーダの全隠れ状態，$\boldsymbol{h}_t$は求めたい decoder の隠れ状態の値である．$\overline{\boldsymbol{h}_s}$と$\boldsymbol{h}_t$より，$\boldsymbol{a}_t$を求められ，$\boldsymbol{a}_t = softmax(score(\boldsymbol{h}_t, \boldsymbol{h}_s^{\top}))$である．ただし，$score()$は一般的に内積である．$\boldsymbol{a}_t$と$\overline{\boldsymbol{h}_s}$の内積を計算して，その総和が context vector $\boldsymbol{c}_t$となる．最後に，$\widetilde{\boldsymbol{h}_t} = tanh(\boldsymbol{W}_c[\boldsymbol{c}_t; \boldsymbol{h}_t])$を計算し，$\widetilde{\boldsymbol{h}_t}$がデコーダからの出力になる．

self-attention[6]は双方向 LSTM と Attention が用いられている分類モデルである．双方向 LSTM は順方 LSTM と逆方向 LSTM からなり，それぞれの方向からの計算によって得られた隠れ状態の値を連結し，連結したものを新たな隠れ状態として用いる．これにより，双方向 LSTM は学習および分類を行う時に将来の単語に関する情報を加えることができる．

図 2 に self-attention による分類モデルを示す．self-attention では，まず順方向 LSTM および逆方向 LSTM の隠れ状態$H = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n)$を計算する．次に，attention の重み$A = softmax(W_{s2}\, tanh(W_{s1} H^{\top}))$を計算する．このとき，$W_{s2}{:}(r \times d_a)$および$W_{s1}{:}(d_a \times 2u)$は重み行列，$u$は隠れ状態の次元数，$r, d_a$はハイパーパラメータである．ハイパーパラメータは文献[4]ではr=30, d_a=350 としている．次に埋め込み行列 M=AHを求めた後，2 度線形結合を行い，クラスごとの出現確率を求める．

2.2. 深層学習の判断根拠の解釈性

DNN の判断根拠を示す手法として，顕著性マップ (Saliency Map) を用いて分類モデルが注目する次元を推定する Vanilla Gradients[7]や SmoothGrad[8]がある．Vanilla Gradients は CNN の入力値に対する出力値の勾配を計算することで，入力画像における分類に大きく寄与する画素を可視化する手法である．SmoothGrad は入力画像にガウシアンノイズを加えて複数のサンプルを作成することで，入力次元ごとの勾配値を計算し平均する．これにより，Vanilla Gradient より分類に重要な画素をより明瞭にハイライトすることが可能となる．これらの文献においては画像に対して検証されており，テキストの主観情報抽出などに対しては検証されていない．

CNN の判断根拠を示す手法のひとつに Grad-CAM[9]がある．これは，CNN の入力値に対する畳み込み層の最終層の勾配値を用いて分類に寄与する画素を可視化する手法である．本稿ではニューロンのみのシンプルな DNN もしくは LSTM を用いて分類を行っているため，Grad-CAM ではなく，SmoothGrad を用いて勾配計算を行った．

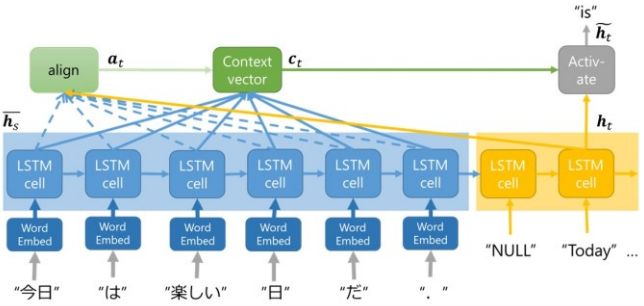


図 1 Attention 付き Encoder-Decoder

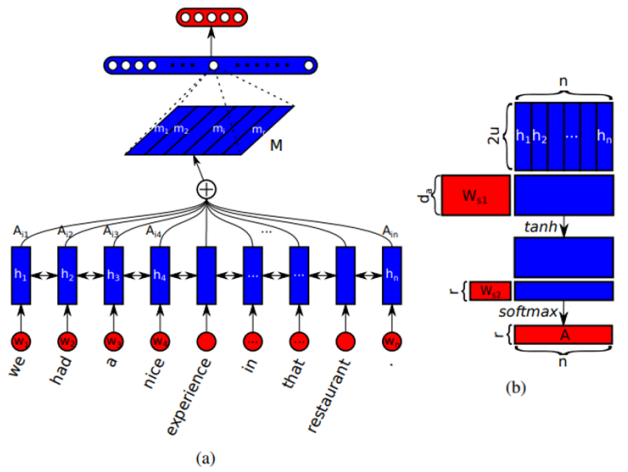


図 2 A sample model structure showing the sentence embedding model combined with a fully connected and softmax layer for sentiment analysis (a)[6].

文献[2]において，LIME という分類モデルの決定を解釈可能にする手法が提案されている．同文献において著者らは機械学習モデルの多くがブラックボックスなモデルであることを指摘しており，機械学習モデルの決定理由の理解が重要であると述べている．また，著者らは画像の背景が雪であるか否かで写真内の動物が狼であるかシベリアンハスキーであるかを判断する学習モデルなどを ”Bad model” と主張している．

文献[10]において，SVM の重みベクトルの絶対値に着目し，SVM の判断に解釈性を付与する手法が提案されている．

3. 文書の学習および分類

本章にて機械学習による主観文書の分類を行い，分類結果を対象に 4 章にて解釈性付与結果を示し，これについて考察を行う．

文献[3]では，形態素解析された文書を Bag-of-Words (BoW) ベクトル表現に変換し，BoW ベクトル表現を SVM および DNN の入力データとして，学習および分

類を行った．そして SVM および DNN の分類の判定理由を提示し，その考察を行った．本稿では文献[3]にある SVM と DNN による分類モデルに加え self-attention による学習と分類を行いその判定理由を提示し，これについて考察を行う．

3.1. 実験環境

まず，大手ショッピングサイト Amazon からレビュー文書を取得した．本実験に使用したレビュー文書のジャンルは書籍および DVD であり，カテゴリ上位 30 商品のレビュー文書である．レビュー文書には 1 から 5 までの値（評価値）が記載されており，本実験では評価値が 1, 2 のものを低評価，評価値が 5 のものを高評価として使用した．これは，評価値が 5 のレビュー文書数に比べて，評価値が 1 のレビュー文書数が大きく下回っていて，評価値 1 のレビュー文書数だけでは学習および分類に使用するデータが少なくなってしまうからである．また，高評価レビューと低評価レビューの件数を同一にするために，30 商品のレビュー群から無作為にレビューを抽出し，これらを学習および分類データとした．実験に用いたレビュー文書の情報を表 1 に示す．表 1 中のレビュー数の列にある P と N はそれぞれ肯定的レビューと否定的レビューの数を表している．

次に，レビューを形態素解析した．形態素解析には MeCab のバージョン 0.996 を使用し，MeCab の辞書には NEologd を使用した．各ジャンルの総形態素数は表 1 のとおりである．

文献[3]の SVM と DNN では，単語のベクトル表現に BoW を用いている．形態素解析によって得られた形態素群と各レビューにおける各形態素の出現頻度から，BoW ベクトルを求めた．ただし，得られた形態素のうち，助詞，記号，名詞数字は除外して BoW ベクトルを求めた．SVM および DNN のハイパーパラメータを表 2 および表 3 に示す．

attention が無い LSTM（ナイーブ LSTM）と self-attention による学習と分類では，単語のベクトル表現に fastText を用いた．fastText のモデルにはウィキペディア日本語コーパスで事前学習したモデルを使用した．本稿ではナイーブ LSTM と self-attention の学習および分類時には品詞の除外は行っていない．ナイーブ LSTM と self-attention のハイパーパラメータを表 4 に示す．3.2 節にてナイーブ LSTM についての正解率も掲載したため，表 4 にはナイーブ LSTM のハイパーパラメータについても掲載した．

学習および分類は書籍ジャンルレビューで学習して書籍ジャンルレビューで分類するもの（同一ジャンル学習分類）と書籍ジャンルレビューで学習し，DVD ジャンルレビューで分類するもの（異ジャンル学習分類）を行った．学習データと訓練データの比率は，同一ジャンルでは書籍ジャンル 80%：書籍ジャンル 20%，異ジャンルでは書籍ジャンル 100%：DVD ジャンル 100%とした．

表 1 実験に使用したレビュー文書の情報

ジャンル	商品数	レビュー数	総形態素数
書籍	30	6000 (P:3000, N:3000)	31835
DVD	30	6000 (P:3000, N:3000)	40324

表 2 SVM のハイパーパラメータ

SVM	
使用ツール	バージョン
SVM-light	6.02
カーネル関数	Liner

表 3 DNN のハイパーパラメータ

DNN		
フレームワーク	Tensorflow 1.7.0	
隠れ層	2	
隠れ層ユニット数	1層目: 256, 2層目: 32	
活性化関数	隠れ層	ReLu
	出力層	Sigmoid
最適化関数	Adam	
学習率	5.00E-04	
損失関数	Cross Entropy	

表 4 LSTM を使用したモデルのハイパーパラメータ

	LSTM	self-attention
フレームワーク	Pytorch 1.3.1	
バッチサイズ	128	
LSTM隠れ層	1	
隠れ層ユニット数	256	
単語ベクトル次元数	300	
最適化関数	adam	
学習率	1.00E-04	
損失関数	Cross Entropy	
r	-	1
d_a	-	350

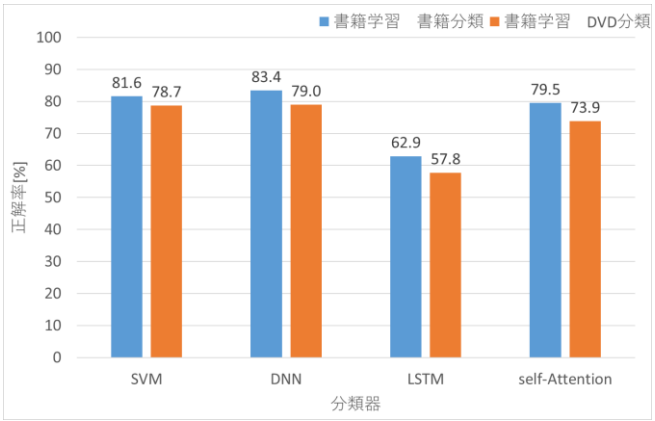


図 3 各分類器による分類精度

3.2. 分類結果

図 3 に各分類器の分類精度を示す．図 3 より，全ての分類器で同一ジャンル分類の精度が異ジャンル分類精度よりも高いことがわかる．また，単純な分類器で

ある SVM と DNN の分類精度が LSTM を用いている分類器よりも高精度であり，特に，ナイーブ LSTM で分類時には高い分類精度が得られなかったことがわかる．self-attention であっても，SVM や DNN のほうが高精度であった．LSTM を用いた分類器では Attention を適用することで正解率が向上することがわかる．

4. 既存手法による分類モデルへの解釈性付与

本章では，SVM および DNN の判断結果に判断根拠の解釈性を与える手法[3]を用いてレビュー文書分類に解釈性を付与した例を提示する．提示するレビューは我々が直観に反すると考えた，異ジャンル学習分類（書籍レビューで学習し DVD レビューの分類）に解釈性を付与した例を示し，これの解釈性について考察を行う．

4.1. 既存手法による分類モデルにおける重要語

本節では分類モデルにおける重要語[3]について述べる．SVM と DNN では文書を Bag-of-Words ベクトルに変換し，そのベクトルを用いて学習と分類を行うことを前提としている．SVM における重要語は重みベクトルの絶対値が大きな次元に対応する単語とした．DNN における重要語は SmoothGrad[8]に基づき BoW の各次元の入力値で出力値を微分し，勾配の絶対値が大きな次元に対応する単語とした．SVM も DNN も正方向に大きいほど高評価に特有の語で，負方向に大きいほど低評価に特有の語となる．

4.2. 既存手法による解釈性付与手法

SVM および DNN では文献[3]を基にして解釈性付与を行った．

SVM は BoW の全次元の線形和が閾値を超えるか否かで判断をするため，次元数が高い時は人間には判断根拠の理解が困難になる．そこで，レビュー文に含まれるうち上位 N 語，下位 N 語の重要語のみの次元の線形和により判断をする簡易化した判断過程を示し，これを判断理由としてユーザに提示する手法[3]を用いる．なお提示する重要語量 N は SVM においては N=200，DNN においては N=3200 とした．

DNN は SVM の様に線形和で判断を行っていないが，本稿では DNN の判断根拠の解釈性の付与の初期段階の研究として，DNN においても SVM 同様に勾配の絶対値の大きな重要語のみの線形和を提示することにより判断理由を示す手法[3]を用いる．

簡易化した判断過程を示す際に，SVM では正方向(褒め)の語は青い文字で，負方向(貶し)の語は赤い文字で示し，DNN では正方向の語は青いマーカで，負方向の語は赤いマーカで示す．

4.3. 既存手法による解釈性付与の例

以下に，タイトルと本文が 500 文字以下のレビューで，SVM 評価値(線形和)が最大(最も強く肯定と判断)であり，正解ラベルも肯定的となっている例（レビュー1）に対して解釈性を付与した．紙面に収めるため，レビューは 500 文字以下のものを対象とした．

レビューは以下の文章である．先頭の 5 は商品の評価値(星の数)である．これの BoW を SVM では上位 100 語，下位 100 語のみで求め，DNN では上位 1600 語，下位 1600 語のみで求めるとレビュー下の式の様になる．レビューの下にある 1 つ目の式が SVM の重要語のみで計算した SVM の評価値で，2 つ目の式が DNN の重要語のみで計算した DNN の評価値となっている．

5　いつ見ても感動する映画です。
1967 年か 1968 年に見た時も感動しました。sound track のレコードも買いました。歌詞も一部覚えました。そして、先日 TV で見て、ぜひ、DVD を自分で持っていたいと思いました。映画の中の音楽は、何度聞いても美しく、しばらくは、耳の中で響いています。

「まし」* 4 + 「中」* 2 +「そして」* 1 + 「年」* 2 +「い」* 2 +「ぜひ」* 1+ 「たい」* 1
= 0.10626 * 4 + 0.14176 * 2 + 0.19965 * 1 + 0.10665 * 2 + 0.09029 * 2 + 0.10884 *1+ 0.10262 * 1 = 1.51355

「度」 * 1 + 「ぜひ」 * 1 + 「TV」 * 1 + 「まし」 * 4 + -「た」 * 5 + 「響い」 * 1 + 「美しく」 * 1
= 0.00017 * 1 + 0.00039 * 1 + 0.00017 * 1 + 0.00016 * 4 + -0.00016 * 5 + 0.00019 * 1 + 0.00019 * 1 = 0.00095

閾値は-0.32437 である．評価値の総和は閾値以上となり，SVM は本レビューを高評価と判断したこととなる．SVM はレビューの判断を行う上で，丁寧語の「まし」などにより高評価と判断したことを理解できる．また判断時の「感動」の評価値は-0.00376，「美し」の評価値は 0.01101 であり，大きな判断根拠とはなっていない．また，全語による評価値の線形和は 1.97046 であり，簡易化しても近い結果が得られている．

DNN では助動詞の「まし」や「た」が根拠として提示されており，判断への影響は大きい．一方で「感動」の値は 0.00007 となっているため，判断する上で大きな判断根拠となっていない．また，「響い」や「美しく」といったジャンルを問わない褒め語も提示している事がわかった．

次に，正解ラベルは低評価であるが，SVM が高評価と判断されたものの中で，最も SVM の評価値が高い(誤っていたもののうち，最も強く肯定と判断)例（レビュー2）の判断根拠を示す．

1　テレビ世代に送る最高のエンタテイメント
特撮は素晴らしいキャストすごい監督職人技テレビ世代に送る最高のエンタテイメント

「すごい」* 1 + 「テレビ」* 2 + 「最高」* 2 + 「素晴らしい」* 1
= 0.10749 * 1 + 0.09814 * 2 + 0.09983 * 2 + 0.17477 * 1= 0.6782

「テレビ」 * 2 + -「キャスト」 * 1 + 「すごい」 * 1 + 「素晴らしい」 * 1
= 0.00016 * 2 + -0.00019 * 1 + 0.00016 * 1 + 0.00021 * 1= 0.0005

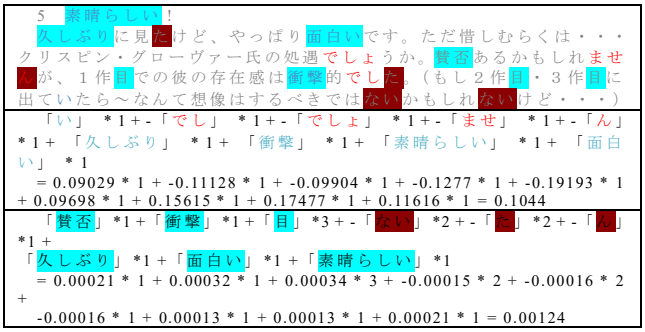
閾値は-0.32437 であり，SVM はこのレビューを高評価と判断したこととなる．全語の評価値線形和は 0.70619 であり，近い値となっている．

SVM は「すごい」，「最高」などを根拠として提示している．このことから，著者の主観では SVM の分類結果には納得がいくものであると考えられる．

DNN もこのレビューの判定を間違えているが，SVM

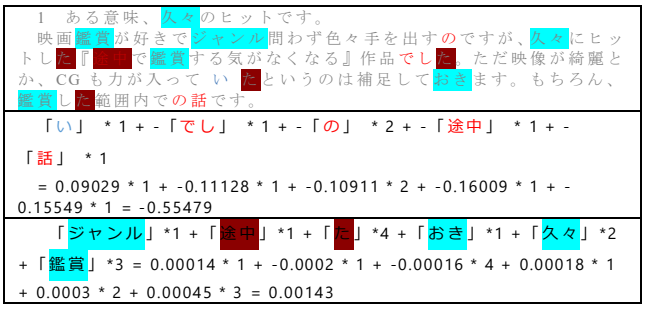
とほとんど同じような語を根拠と提示している．このことから，SVM と同様に DNN の分類結果も納得がいく結果であると考えられる．また，示された根拠も多くの人間に納得できるものであると予想をしている．

次に，DNN による分類で，勾配と出現回数の内積(線形和)が最も大きく，分類が成功した例（レビュー3）を示す．



SVM と DNN は「衝撃」，「素晴らしい」，「面白い」のような，同一の語を判断根拠として提示している事が分かる．SVM はジャンルを問わない褒め語である「素晴らしい」,「面白い」を根拠として提示している．しかし，その一方で提示した語の半分程度を「でし」，「でしょ」といった助動詞が占めており，DNN は SVM より助動詞の提示量が少ないことが分かる．また，DNN では「ない」が重要語となっていることから，否定文であれば貶す文章であるとみなすような分類モデルを構築している可能性が考えられる．

次に，正解ラベルは低評価であるが，DNN が高評価と判断したものの中で，勾配と出現回数の内積(線形和)が最も大きく，分類が失敗した例（レビュー4）の判断根拠を示す．



SVM においては，ジャンルを問わない貶し語を重要とみなして判断していないことから，提示された語には納得感が無いことが考えられる．DNN も助動詞の「た」や，「久々」，「鑑賞」などの納得感の無い語を選択していた．なお,「好き」は評価値 0.00004,「ヒット」は-0.00005,「綺麗」0.00002 であった．全語の勾配値の和は 0.00220 で簡易化しても近い結果が得られた．

本レビューは，商品を貶す文章であるが故意に肯定的な表現を多用する表現方法を用いているため，文章を BoW ベクトルに変換したために，SVM や DNN はレビューを高評価と判断したと考えられる．

4.4. 既存手法による解釈性付与についての考察

文献[3]や 4.3 節の解釈性付与から，SVM は丁寧語を用いている丁寧な文であれば高評価であり，否定文や貶す語「ない」ではなく文意を反転させる語を含む文が低評価であることを活用して判断していることが分かる．これは，人間の主観にて本質的でない特徴を活用していると考えられ,高い精度が得られたとしても，背景を根拠に動物を予想するのに類似した Bad Model の側面を持つとも考えられる．

DNN は SVM と異なり，「鑑賞」,「途中」などの学習時のデータに特有な語を判断根拠としていることが確認できた.このことから,SVM とは異なるが Bad Model を構築していると考えられる．しかし，文意を反転させる「ない」や丁寧語での判断よりも学習時データに特有な語を判断根拠としている方が本質的な特徴であると考える．このため，DNN は SVM より人間の主観での判断に近いモデルを構築していると考えられる．

SVM と DNN では，どちらも正解率 8 割程度を達成しているものの，どちらのモデルも Bad Model の側面が大きいと考えられる．また，SVM では紹介した他にも「丁寧な文であれば好評価」,「否定文であれば低評価」と判断している例が多く確認された．

5. 提案手法による解釈性付与

本章にて self-attention を用いた文書の分類を行い，それに対する解釈性の付与手法を提案する．

5.1. self-attention における重要語

self-attention における，時刻 t における Attention の重み$A = softmax(W_{s2}\ tanh(W_{s1}H^{\mathrm{T}}))$を，時刻 t における単語の重要語とする．

5.2. 提案手法による解釈性付与手法

self-attention モデルにおける重要語を示す．提案手法における重要語は，既存手法とは異なり，どの語の時点での LSTM の出力を重要視するかを示すものである．この重要語は Attention の重みの値が大きいものほど，濃く赤い背景，小さくなるほど（0 に近いほど）薄く赤い背景で示す．値の差異が大きくなるように各単語の重要語を 5 乗した値を背景色とする．

6. 提案手法による解釈性付与

6.1. 解釈性付与の例

Attention の重みを基にして LSTM の分類結果に解釈性を付与した例を提示する．提示するレビューは 4 章と同様に，我々が直観に反すると考えた，異ジャンル学習分類（書籍レビューで学習し DVD レビューの分類）に解釈性を付与した例を示し，これの解釈性について考察を行う．

4.3 節のレビュー1 を self-attention によって分類し，解釈性を付与した結果を図 4 に示す．

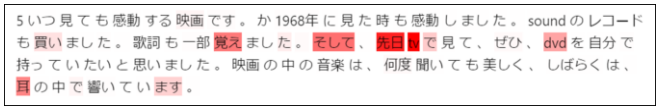


図 4 self-attention によるレビュー1 への解釈性付与結果

self-attention はレビュー1 を肯定的文書であると判断し正しく分類した．表 5 に attention 重みが特に大きかった上位の 10 語を示す．

表 5 図 4 の文書における attention の重み上位 10 語

	tv	先日	そして	覚え	耳	dvd	で	ます	買い	た
attention weight	0.05949	0.05784	0.05302	0.05224	0.05126	0.04826	0.04319	0.04200	0.03863	0.03449

self-attention の attention 重み上位 10 語には，「ます」，「そして」などの文章の切れ目の単語の attention 重みが大きく，特に色が濃い部分の文章の切れ目の時点で LSTM からの出力を大きくし，結論を出そうとしていることがわかる．また，ジャンルを問わない褒め語であると考えられる「感動」（0.016729，0.005065）や，「美しく」（0.000408）などの単語が出現した時点では大きく判断を下していないことがわかる．

次に，4.3 節のレビュー2 を self-attention によって分類し，解釈性を付与した結果を図 5 に示す．

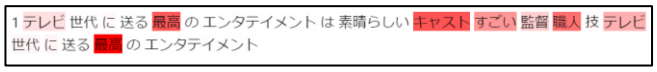


図 5 self-attention によるレビュー2 への解釈性付与結果

self-attention はレビュー2 が肯定的文書であると判断し誤って分類した．表 6 に attention 重みが特に大きかった上位の 10 語を示す．

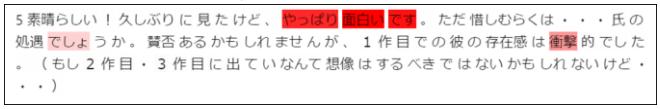
表 6 図 5 の文書における attention の重み上位 10 語

	最高	キャスト	最高	職人	すごい	テレビ	監督	テレビ	に	世代
attention weight	0.09869	0.09111	0.09053	0.08777	0.08266	0.07800	0.07034	0.06509	0.05144	0.04836

表 6 の文書中の単語の attention 重み上位 10 語から，self-attention では，「最高」，「職人」，「すごい」が登場する時点での LSTM の出力が重要であることがわかる．ジャンルを問わない褒め語であると考えられる「素晴らしい」の値は，0.030539 であり，この時点での LSTM の出力は重要視していないことがわかる．

次に，4.3 節のレビュー3 を self-attention によって分類し，解釈性を付与した結果を図 6 に示す．

図 6 self-attentionx によるレビュー3 への解釈性付与結果



self-attention はレビュー3 が肯定的文書であると判断し正しく分類した．表 7 に attention 重みが特に大きかった上位の 10 語を示す．

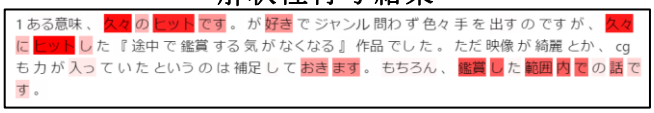
表 7 図 6 における attention の重み上位 10 語

	面白い	です	やっぱり	衝撃	でしょ	3	2	目	的	<unk>（クリスピン・グローヴァー）
attention weight	0.08824	0.08698	0.08535	0.07382	0.06235	0.03997	0.03510	0.03502	0.03380	0.03283

表 7 より，「面白い」の attention 重みが最も大きく，周辺にある「やっぱり」や「です」が大きくこの箇所もとに判断しているため，肯定的レビューであると判断したと考える．また，「素晴らしい」の重みは小さく（0.011775），大きな判断根拠にはならない．

最後に，4.3 節のレビュー4 の分類と解釈性付与結果を図 7 に示す．

図 7 self-attention によるレビュー4 への解釈性付与結果



self-attention はレビュー4 が肯定的文書であると判断し誤って分類した．表 8 に attention の重みが特に大きかった上位の 10 語を示す．

表 8 図 7 の文書における attention の重み上位 10 語

	ヒット	久々	久々	ヒット	鑑賞	し	内	範囲	で	ます
attention weight	0.04570	0.04525	0.04472	0.04438	0.04257	0.04174	0.04168	0.04167	0.04136	0.03871

表 8 の文書中の単語の attention 重み上位 10 語には，「ヒット」があり，この箇所を基に肯定的と強くみなしている事が考えられる．重みが大きい語から，上位 10 語ではないものの，「好き」（0.038303）の値が大きいことがわかる．「綺麗」の値は 0.000394 で小さかった．「途中で鑑賞する気がなくなる」という否定的レビューであると人間は考えると思われる箇所の attention 重みは小さく，分類する上で重要でないとみなされ，分類を誤ってしまった可能性が考えられる．レビュー4 は皮肉を含む文章であり，self-attention も低精度のモデルのため，分類を誤ってしまった事も考えられる．

6.2. 提案手法による解釈性付与についての考察

SVM や DNN の解釈性付与と比較すると，語が登場する箇所によって，同じ語でも attention 重みは違うことが分かり，場所によって重要度が変わることが考えられる．

self-attention の attention 重みが大きい箇所は，文頭や文末が多く，それら周辺の attention 重みが大きかったことから，文中の語よりも文頭や文末の時点で判断を下しており，文中の語に判断を下す上で重要な語が含まれている場合にそのような語を考慮しないような分類モデルになっており，結果として分類精度の低下を招いているのではないかと考えられる．

7. おわりに

本稿では，SVM および DNN による文書分類モデルと self-attention による文書分類モデルで学習および分類し，モデルの解釈を試みた．そして，解釈結果を示し有効性について考察をした．今後は，大きく異なる分野同士の分類や Transformer[11]による分類器の解釈性付与についての考察を行う予定である．

謝辞

本研究は，JSPS 科研費 15H02696，17K00109，18K11277 の助成を受けたものである．

本研究は，JST，CREST JPMJCR1503 の支援を受けたものである．

参 考 文 献

[1] Grégoire Montavon, Wojciech Samek and Klaus-Robert Müller, Methods for Interpreting and Understanding Deep Neural Networks, Digital Signal Processing Volume 73, Pages 1-15, February 2018.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). ACM, New York, NY, USA, 1135-1144. DOI: https://doi.org/10.1145/2939672.2939778

[3] 中村 鴻介，山口 実靖，" 機械学習による主観文書分類結果の解釈性の付与に関する一考察 ", WebDB Forum 2019 論文集，Vol. 2019, pp. 17-20

[4] Sepp Hochreiter and Jürgen Schmidhuber, “Long short-term memory”. Neural Computation 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735., 1997.

[5] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, The International Conference on Learning Representations (ICLR ’14), 2014.

[6] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou and Yoshua Bengio, A Structured Self-attentive Sentence Embedding, The International Conference on Learning Representations (ICLR ‘17), 2017.

[7] Karen Simonyan, Andrea Vedaldi and Andrew Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Workshop on ICLR, 2014.

[8] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas and Martin Wattenberg, SmoothGrad: removing noise by adding noise, Workshop on Visualization for Deep Learning in ICML, 2017

[9] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam and Devi Parikh Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626

[10] S. Shirataki and S. Yamaguchi, "A study on interpretability of decision of machine learning," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 4830-4831 doi: 10.1109/BigData.2017.8258557

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin, “Attention Is All You Need”, NeurIPS, 2017