4Q-07

ノイズを用いた深層学習における学習モデルの解釈性に関する一考察

中村鴻介[†1]　佐藤悠祐[†2]　山口実靖[†1†2]
[†1] 工学院大学　[†2] 工学院大学大学院

1. はじめに

近年，ディープニューラルネットワーク(DNN)，畳み込みニューラルネットワーク(CNN)，リカレントニューラルネットワーク(RNN)といった深層学習(Deep Learning)が人工知能の実現手法の一つとして普及しつつある．深層学習は従来の機械学習手法に比べて，画像認識，言語認識，分類問題などの様々な問題に対し高精度な推論ができると期待できる学習手法であり，様々な分野での応用がされている．しかし，深層学習には推論結果についての説明性がないものが多いという問題点[1] [2] が指摘されている．たとえば裁判における判決など説明性の付与が必須の状況や，学習の結果に基づき責任のある決定をする場合などに推論の解釈性が提供されている方が好ましい状況など，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状況は多いと考えられる．

また一方で，オンラインショッピングサイトなどの普及により，ある対象に関する主観的な評価情報が多数存在するようになった．評価の対象にはショッピングサイトの商品等があり，主観的評価情報にはレビュー等がある．これらの評価情報を解析し，肯定的なものと否定的なものに分類して提示することは，提供者や利用者の双方にとって有益であると考えられる．また，主観文書の内容に基づく肯定的，否定的の自動的な分類手法の確立は重要であると考えられる．

本研究では，SVM および DNN を用いてレビュー情報を高評価であるか低評価であるかの分類を行う．そして，その分類の判断根拠を提示することに取り組む．具体的には，レビュー文書を Bag-of-Words(BoW)ベクトルに変換してこれを説明変数とし，文書が高評価であるか低評価であるかを目的変数として，SVM および DNN にて学習と分類を行う．そして，SVM においては重みベクトルの重みの絶対値が大きな次元に対応する単語(形態素)が重要であると考え，この重要語を判断根拠の一つとして提示する．DNN においては，ノイズを用いて解釈性を得る既存手法 SmoothGrad [3] を用いて説明変数による目的変数の微分を求め，これの絶対値が大きい語を重要語とみなし，これの提示を行う．本研究では，DNN の解釈性の研究の初期段階として，一次式に基づく解釈性の考察を行う．

2. 関連研究

DNN の判断根拠を示す手法として，Vanilla Gradient [4] や SmoothGrad [3] がある．Vanilla Gradient 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顕著にハイライトすることが可能となる．これらの手法は画像に対してのみ検証されており，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．

Ribeiro らは，機械学習モデルの多くがブラックボックスであることを指摘し，決定理由の理解が重要であると主張している．また，LIME という決定を解釈可能にする手法を提案し[2] ，画像の背景が雪であるか否かで狼の写真であるかシベリアンハスキーの写真であるかを判断する学習モデルなどを悪いモデルと主張している．

文献[5] 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. レビュー分類

ショッピングサイト Amazon からレビュー文を抽出し，その文の BoW ベクトルを SVM および DNN を用いて以下の様に学習および分類を行った．

まず，ビール，書籍，DVD の 3 ジャンルより，レビュー数上位 30 商品のレビューを取得した．レビューには評価値(1~5)が記載されており，本研究では，評価 1, 2 を低評価，評価 5 を高評価とした．これは，評価 5 のレビュー数と比べ，評価 1 のレビュー数が大きく下回っていたためである．また，低評価レビュー数と高評価レビュー数が同一になるように，30 商品のレビュー群から，無作為にレビューをピックアップし，これらを学習と分類用のデータとした．ピックアップしたレビュー数は高低評価それぞれビールにおいて 200，書籍において 3000，DVD において 2924 である．

次に，ピックアップしたレビュー群を形態素解析機 MeCab を用いて形態素解析し，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，それを SVM および DNN に入力し，学習と分類を行った．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．また，学習と分類に用いた BoW ベクトルの次元数(すなわち形態素の総種類数)はビールにおいて 6944，書籍において 28687，DVD において 40324 である．学習と分類に用いたデータの比率は SVM, DNN ともに学習 : 分類 = 8 : 2 である．

表 1 分類精度（%）

レビュー群	SVM	DNN
ビール	71.79	77.50
書籍	81.59	83.41
DVD	86.13	86.67

A Study on Interpretability of Learning Model of Deep Learning
[†1] Kosuke Nakamura, Saneyasu Yamaguchi, Kogakuin University
[†2] Yusuke Sato, Kogakuin University Graduate School

表 2 SVM 正方向に重みが大きい形態素

ビール		書籍		DVD	
重み	形態素	重み	形態素	重み	形態素
0.1554	美味しい	0.2060	マスコミ	0.3993	最高
0.1506	飲み	0.1895	くれ	0.2693	満足
0.1367	作品	0.1659	とても	0.2530	素晴らしい
0.1262	時	0.1638	そして	0.2284	度
0.1262	ある	0.1636	られる	0.1939	時代
0.1230	ます	0.1620	素晴らしい	0.1918	ジョーカー
0.1125	夏	0.1615	くれる	0.1806	良かっ
0.1075	苦み	0.1610	今	0.1744	心
0.1044	し	0.1610	一気に	0.1729	世界
0.1025	手	0.1521	いく	0.1713	とても

表 3 SVM 負方向に重みが大きい形態素

ビール		書籍		DVD	
重み	形態素	重み	形態素	重み	形態素
-0.2524	ない	-0.2683	なら	-0.4446	残念
-0.2141	な	-0.2616	残念	-0.2660	駄作
-0.2136	味	-0.2225	芥川賞	-0.2546	こんな
-0.1940	た	-0.2004	ない	-0.2525	がっかり
-0.1932	期限	-0.1981	正直	-0.2372	すぎ
-0.1791	なかっ	-0.1972	期待	-0.2345	評価
-0.1725	残念	-0.1890	すぎ	-0.2180	ん
-0.1602	方	-0.1844	つまらない	-0.2064	最低
-0.1424	の	-0.1722	ん	-0.1927	正直
-0.1393	賞味	-0.1721	方	-0.1865	期待

表 4 DNN 正方向に傾きが大きい形態素

ビール		書籍		DVD	
勾配値	形態素	勾配値	形態素	勾配値	形態素
0.009004	ベルギー	0.000697	安く	0.000798	保田
0.008977	すっきり	0.000660	町	0.000738	真夏
0.008956	うまい	0.000620	かつ	0.000730	最優秀
0.008556	美味し	0.000615	デフォルメ	0.000705	御社
0.008497	爆発	0.000592	来	0.000687	ダンスフォーメーション
0.008232	良かっ	0.000585	なんなく	0.000684	里山
0.007611	冬	0.000583	胸	0.000682	幸運
0.007589	持ち帰り	0.000579	やっぱし	0.000679	フィーチャリング
0.007561	差し入れ	0.000574	罰する	0.000673	見失わ
0.007458	定番	0.000573	しずか	0.000661	うれしかっ

表 5 DNN 負方向に重みが大きい形態素

ビール		書籍		DVD	
勾配値	形態素	勾配値	形態素	勾配値	形態素
-0.011904	水っぽい	-0.000806	今年度	-0.000765	がっくり
-0.009764	第	-0.000686	寒気	-0.000724	インターネット
-0.009000	薄い	-0.000661	ダイナミック	-0.000688	加護
-0.008064	くさい	-0.000660	そんなに	-0.000675	最
-0.008032	手違い	-0.000650	そもそも	-0.000670	たまらなく
-0.007741	つい	-0.000649	陳腐	-0.000662	こみ上げる
-0.007722	着く	-0.000603	すじっ	-0.000661	ギブアップ
-0.007423	表記	-0.000596	あざとい	-0.000654	はせ
-0.007204	くれ	-0.000572	うーん	-0.000648	困ら
-0.007081	おもう	-0.000567	白け	-0.000648	それなり

DNN は以下の設定で実行した．ネットワークモデルは隠れ層が 2 層で，1 層目のユニット数は 256，2 層目のユニット数は 32 とした．隠れ層の活性化関数に ReLu，出力層の活性化関数にはシグモイド関数を用いた．損失関数に Binary Cross Entropy を用いた．

SVM および DNN での分類精度(正解率)を表 1 に示す．単位は%である．表 1 より，3 ジャンル全てにおいて DNN が，SVM より高い正解率を実現したことがわかる．ビールの分類精度は SVM と DNN ともに，書籍と DVD より低くなっているが，これはビールの学習レビュー数が少ないためであると考えられる．

4. 学習モデルの解析

本章において，学習モデルの解析結果を示す．まず，SVM の学習モデルの重みベクトルにおける正負の重み上位 10 個の次元を表 2 と表 3 に示す．重みベクトルの各要素は，BoW の各形態素と対応しており，重みの絶対値が大きい形態素は SVM が肯定的か否定的かを判断する上で重要とみなした語と考えることができる．表 2 は重みがプラス方向に大きい形態素で，表 3 は重みがマイナス方向に大きい形態素である．また， DNN の学習モデルにおいて，SmoothGrad により求めた正負の勾配値上位 10 個を表 4 と表 5 に示す．表 4 は入力値に対する出力値の勾配がプラス方向に大きい形態素で，表 5 は入力値に対する出力値の勾配がマイナス方向に大きい形態素である．

表 2～表 5 から，SVM と DNN の正と負の方向の重要語には人間の主観において高評価や低評価を表現すると考えられる形態素(美味しい，うまい，残念，くさいなど)が多く含まれていることを確認できる．また，そうとは考えられない語(評価，インターネット)も含まれていることも分かる．我々は，SVM の重要語の方が人間の主観において高評価や低評価と考えられる語が高い割合で含まれていると考えた．この考えに基づくと，今回の事例では DNN の方が高い正解率を達成しているが適切さの低いモデルを構築していることを予想することも可能となる．

5. おわりに

本稿では，SVM と DNN による分類の判断根拠の解釈性に着目し，重みベクトルや勾配に基づき主観情報を含むレビュー文書群の分類に解釈性を付与することを行った．

今後は，異なるジャンル間での分類の解釈性の考察，DNN の隠れ層の入出力値と出力層の出力値の関係の考察，非線形の式に基づく解釈性の考察などを行っていく予定である．

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

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

[5] 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