6L-08

アテンションを用いた深層学習の分類結果に対する解釈性付与に関する一考察

中村　鴻介†　山口　実靖†
†工学院大学大学院

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合させて構成されており，解決したいタスクに応じて深層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)，再帰ニューラルネットワーク(RNN)やLSTMなどが用いられている．これらは従来の機械学習手法に比べて，自然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，幅広く応用されている．しかし，文献[1] [2] などで，深層学習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．裁判の判決や，経営判断の株主への説明や政治家の政治判断の有権者への説明など，判断結果に対して責任が求められる状況は多く，解釈性や説明性の付与は重要であると考えられる．

本稿ではまず，self-attentionによる2つのニュース記事の分類モデルを構築し，これを用いて記事の分類を行う．次に，既存手法[3] をナイーブに拡張した手法(以下，SG絶対値と呼ぶ)とAttention値を用いる手法に着目し，解釈性付与を行い性能の評価を行う．

2　関連研究

self-attention[4] [5] 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 は順方向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 LSTM の出力値を連結し，連結したものを双方向 LSTM の出力として用いる．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．

DNN の判断根拠を示す手法の一つに，顕著性マップを用いて分類モデルが注目する次元を推定するSmoothGrad[3] がある．SmoothGradはCNNの入力値にガウシアンノイズを加えて，入力次元ごとの入力値の変化に対する出力値の変化の勾配値を計算し平均することで，入力画像における分類に大きく寄与する画素を抽出する手法である．この手法に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．我々は，過去の文献[7] において DNN によるテキスト分類に対してSmoothGradを適用したが，LSTMやself-attentionを用いたテキスト分類には適用されていない．

文献[2] において，LIME という分類モデルの決定を解釈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を示し，これを “Bad model” と主張している．

文献[6] において，SVM の重みベクトルに着目し，SVM の判断に解釈性を付与する手法が提案されている．

Interpretability of Classification of Deep learning based on Attention
†[1] Kosuke Nakamura, Saneyasu Yamaguchi, Kogakuin University Graduate School

3　判断根拠となる語の抽出手法

Self-attentionによる文書分類に判断根拠を付与する手法としてSG絶対値とAttention値を用いる手法を定義する．

SG 絶対値を用いる手法は，SmoothGrad[3]をナイーブに Self-attention による文書分類に適用した手法であり，入力文(ベクトル表現された単語の配列)に対してノイズを付加し，それによる出力値の変化を調査する．入力語のベクトルの各次元において微分値を求め，この微分値のベクトルの絶対値が大きい語を結論に大きな影響を与えた判断根拠語とする．Attention 値を用いる手法は，Attention値が大きい語を判断根拠語とする．

4　性能評価

本章にて両手法の性能評価を行う．

まず評価で用いる分類について述べる．分類対象にはlivedoor ニュースコーパス 9 ジャンルの中から「家電チャンネル」と「エスマックス」を選択し，これを Self-attention による分類を行った．9 ジャンルのうちのこの 2 ジャンルの分類が最も高正解率(98.6%)であったため本稿の性能評価にて採用した．それぞれ，家電とモバイルガジェットを話題にした記事である．家電チャンネルの記事数は 864，エスマックスの記事数は 870 であった．ニュース記事の形態素解析にはMeCab 0.996を使用し，MeCab辞書には NEologd を使用した．単語のベクトル表現にはfastText を用い，fastText のモデルにはウィキペディア日本語コーパスの事前学習モデルを用いた．品詞などの除外は行っていない．Self-attentionのハイパーパラメータ等は，フレームワーク Pytorch 1.3.1, 双方向 LSTM 出力次元数 512, 単語ベクトル次元数 300, 最適化関数 Adam，学習率 0.001, バッチサイズ 128, 損失関数 CrossEntropy である．訓練データとテストデータの比率は，80% : 20% とした．学習は訓練データの損失関数が収束するまで行った．

self-attention により学習し 2 種類の記事を分類したときの精度(accuracy)は 98.6%であった．テストデータ数は家電チャンネルが 174，エスマックスが 173 であったが，家電チャンネルをエスマックスと誤って分類したものが 1 つ，逆方向に誤って分類したものは 4 つであった．

文の推移における SG 絶対値と Attention 値の推移を図 1 に示す．これは，エスマックス記事群から無作為に抽出した 1 記事の各形態素の SG 絶対値と Attention 値の推移である．当記事の形態素数は 490 であり，横軸が当記事の形態素の登場順である．両手法で類似の結果が得られているが完全には同一ではないことが分かる．値のピークは共に句点の周辺であり，句点前後の文脈が記事分類に

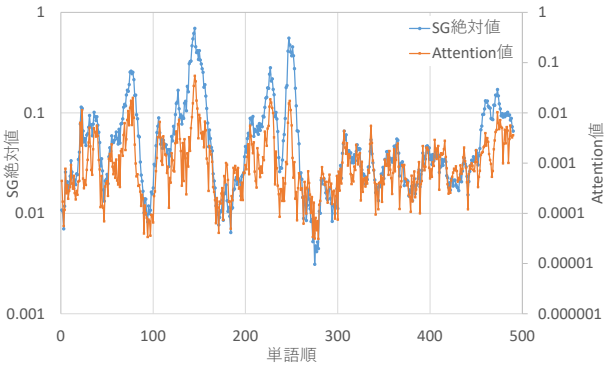

図 1 SG 絶対値と Attention 値の推移

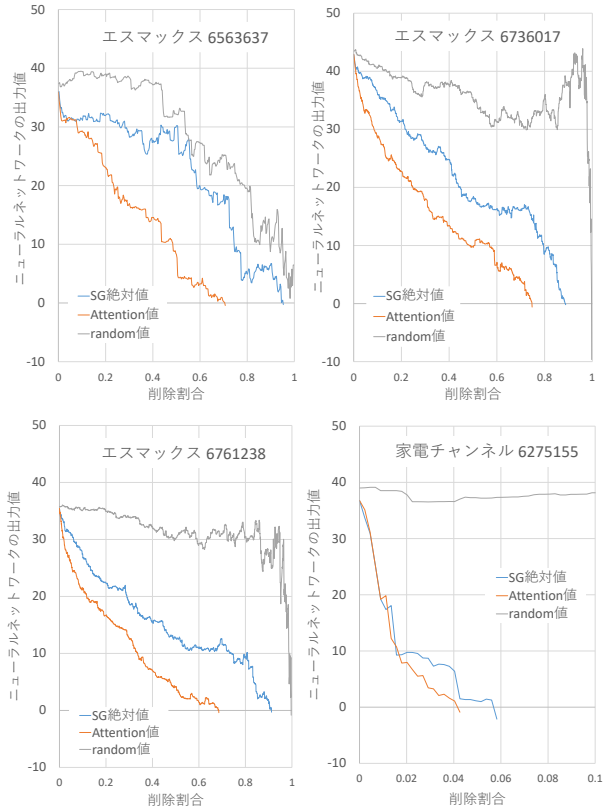

図 2 単語割合とニューラルネットワーク出力値

おいて重要な局面と判断され SG 絶対値や Attention 値が大きくなったと考えられる．

当文書の各語の SG 絶対値と Attention 値の相関について述べる．上記の記事の各語の SG 絶対値の順位と Attention 値の順位の相関係数は 0.762 であり，両手法には強い相関があり，似た語を根拠と見なす傾向があることが分かる．また，完全に同一でもないことも分かる．

次に，両手法が提示した語がどの程度強く判断根拠となっているかを評価する．SG 絶対値もしくは Attention 値が高い順に文書から形態素を削除していき，いくつの形態素を削除したら分類を誤るかの検証を行った．各ジャンルから無作為に 5 つずつの記事を選択し検証対象とした．結果を表 1，表 2 に示す．表 1 より，SG 絶対値よりも Attention 値の上位語を削除する方が早く分類を誤り，入力と出力の関係を調査する SmoothGrad をナイーブに用いるよりも Attention 値に着目した方がより判断根拠を抽出できていることが分かる．表 2 における削除数の－は全形態素を削除し形態素数を 0 にしても判断分類を誤らなかった例である．判断を誤った例は 1 例のみであるが同様に Attention 値に着目した方が判断根拠を抽出できていることが分かる．最後に値が高い順に形態素を削除して行った際の LSTM 後段のニューラルネットワークの出力値の変化を図 2 に示す．対象は表 1 の上の 3 記事と，表 2 の判定を誤った 4 番目の記事である．比較のためにランダム順位単語を削除した場合の出力値の推移も示す．図より Attention 値の手法がより少ない削除数で出力値が大幅に減少しており,特に上位語における大幅な減少が確認できより判断根拠を抽出できていることが分かる．

表 1 エスマックス記事における分類を誤る削除数

エスマックス記事id	SG絶対値削除数	Attention値削除数	記事の形態素
6563637	365	271	382
6736017	1093	920	1229
6761238	1660	1248	1819
6770304	303	173	313
6857034	264	204	490

表 2 家電チャンネル記事における分類を誤る削除数

家電チャンネル記事id	SG絶対値削除数	Attention値削除数	記事の形態素
5990886	-	-	462
6094797	-	-	394
6237932	-	-	297
6275155	27	20	446
6482447	-	-	379

5 おわりに

本稿では，self-attention を用いる深層学習による文書分類の結果の判断根拠の抽出方法について考察を行った．評価の結果，微分により入力値の変化に対する出力値の変化を調査する既存手法より Attention 値を直接用いる手法がより強い根拠を抽出できることが分かった．

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 and Martin Wattenberg, SmoothGrad: removing noise by adding noise, Workshop on Visualization for Deep Learning in ICML, 2017

[4] 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.

[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] 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

[7] 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