特徴マップによる説明の信頼性評価手法の定量的分析

浅妻 佑弥 †1　塙 一晃 ‡†2　乾 健太郎 †‡3
† 東北大学　‡ 理化学研究所
[1]asazuma.yuya.r7@dc.tohoku.ac.jp　[2]kazuaki.hanawa@riken.jp　[3]inui@tohoku.ac.jp

1　はじめに

機械学習モデルの性能が日々向上している一方で，未だに多くのモデルは自身の動作について十分な説明を行うことができない．特に，医療分野においては判断結果に根拠を持つことが要求されているため，機械学習モデルの動作を説明するための手法が研究されている．

説明手法の優劣を議論するには説明の有効性を評価する必要がある．最も単純な方法は人間の直観に基づく方法 (Plausibility) で，人手でアノテーションした特徴マップとの一致度を測る．

しかし，人間の直観に基づく手法のみでは不十分であることが指摘されており [1]，あくまで説明は判断根拠の明示を目的にしているため，機械の内部動作に基づいていることが要求される (Faithfulness)．ゆえに，人間の直観に頼らない評価法も必要になる．

そのため，各研究において信頼性に則した説明手法の評価が行われている [2][3][4][5][6]．多種多様な評価方法で説明手法の信頼性が主張されている一方で，それらの評価方法自体の関係性は十分に考慮されていない．また，データセットやモデルの特性から受ける影響に関しても分析されていない．よって，同一の信頼性を議論しているつもりでもそれぞれ別種の性質を評価している可能性が存在する．

本研究では，既存の信頼性評価手法について定量的な評価を行う．問題の簡略化のために扱う説明の形式を特徴マップに限定し，以下の解明を目的に実験を行う．

- データセットの特性による評価結果の変化
- 複数の評価手法における結果の一貫性・変化

最終的に，各評価手法がどの程度データセットの影響を受けるか分析し，各手法に一貫する性質について考察を行った．

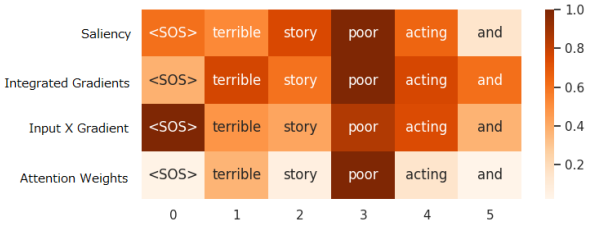

図 1　ある入力例に対する特徴マップの生成例

2　説明の生成手法

1 節で述べたように，本研究では説明として**特徴マップ**のみを扱う．特徴マップとは，出力要素に対する入力要素の重要度を求める手法である．ある感情分析タスクにおける特徴マップの例を図 1 に示す．感情に関わる単語に強い重要度が与えられていることが確認できる．

本節では，これ以降，検証に使用する四つの特徴マップの生成手法について説明する．

2.1　Saliency Map

Saliency Map は入力に対する出力の勾配を求めるシンプルな手法である [7]．勾配の逆伝搬によって容易に計算することができるため，ベースライン手法として扱われることが多い．

2.2　Input X Gradient

上述した Saliency Map を拡張した手法であり，勾配に入力特徴量を乗算する [8]．単純な手法であるが，生成された特徴マップは入力特徴量を考慮することが可能になる．

2.3　Integrated Gradients

Integrated Gradeints は，文献 [9] で考案された特徴マップ生成法である．生成法が満たすべき公理として Sensitivity と Implementation Invariance を定義し，これらを満たす手法としてベースライン入力から入力例に対するモデルの勾配を積分することで，特徴

表 1　本研究で使用するデータセットの統計値

コーパス名	IMDB	SST	AG News	20 News
Train Size	17212	6920	51000	1236
Valid Size	4304	872	9000	310
Test Size	4363	1821	3800	387
単語長 (平均)	181	20	38	171
単語長 (中央)	162	19	37	90
単語長 (90%)	310	31	49	345

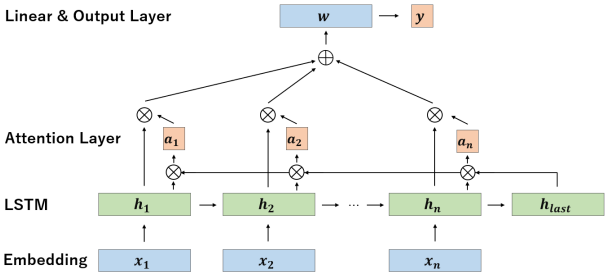

図 2　LSTM + Attention model の図解

マップの生成を行う．

2.4　Attention Weights

近年の NLP 界隈において広く普及した手法に Attention Mechanism が存在する [10]．隠れ層の重み付けを行うため，Attention Weights は入力特徴の重要度に近しいものとして学習される．ゆえに，一種の特徴マップとして扱うことができる [11]．

3　実験条件

3.1　データセット

本稿ではテキスト分類タスクを分析の対象とする．コーパスとして IMDB，SST，AG News，20 News を使用する．ただし問題の単純化のためにすべてのコーパスを 2 値分類タスクに変換する．例として，20 News は 20 値の分類問題であるため，9 番と 10 番ラベルの 2 値分類問題として扱う．

表 1 に使用するデータセットの統計値を示す．この条件設定は文献 [4] に合わせたものである．

3.2　モデル

テキスト分類モデルとして LSTM に基づくモデルを使用する．Attention Weights を取得する必要があるため，LSTM + Attention Layer を含むモデルを使用する．図 2 に使用したモデルを図解する．

上述のモデルについて，3.1 節のコーパスで訓練を行った．訓練後のテストセットにおける Accuracy, F1 スコアを表 2 に示す．

表 2　データセット毎の訓練済みモデルの精度

	IMDB	SST	AG News	20 News
Accuracy	.884	.766	.939	.770
F1 Score	.886	.759	.939	.767

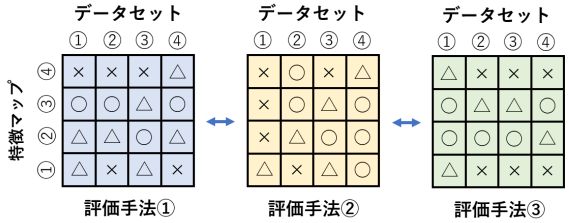

図 3　本研究の実験手順の図解

3.3　評価手法

本研究では図 3 に示すように，説明手法の信頼性を評価する 3 つの手法について検証を行う．4，5，6 節において，タスク・コーパスの違いにおける傾向の変化を評価する．7 節において，手法横断的に分析を行う．

4　ERASER 法を用いた検証

特徴マップの信頼性評価手法として著名な研究に **ERASER** 法がある [3]．これは，特徴マップの降順に入力特徴を削除した際の出力の変化を観測する評価方法である．同様の評価手法が，人間の主観的評価に依存しない評価として多くの研究に採用されている [6][12]．

本実験で使用する ERASER 法は以下の手順で構成される．

1. 出力ラベルが変化するまで，対応する特徴マップの値が高い順に入力要素を MASK する
2. **削除率** = MASK した要素数 / 入力要素数 を計算
3. 全要素を MASK しても変化しなかった場合，削除率 = 1.0 とする

特徴マップが入力要素の重要度を捉えていた場合，重要な入力特徴量から消去することになるため，早い段階で出力が変化すると考えられる．よって，多くの入力における削除率が低いほどその解釈性手法が優秀であると述べることができる．

4.1　実験

3 節で述べた実験設定において ERASER 法の検証を行う．データセット毎の傾向を確認するため，消去率の平均を求めたものを表 3 に示す．

表 3　タスク・解釈手法毎に計算した削除率の平均値

	IMDB	SST	AG News	20 News
Saliency	.570	.508	.677	.610
Integrated Grad	**.095**	**.271**	**.353**	.274
Input X Grad	.101	.292	.397	**.251**
Attention	.479	.545	.644	.532

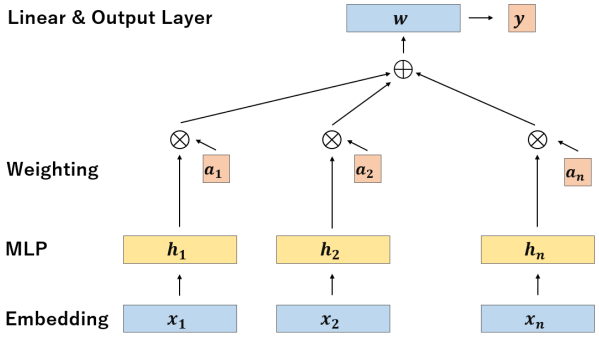

図 4　弱学習器のモデル図解

データセット間で比較を行うと，同じ説明手法でも削除率の値が全く異なることが確認できる．表 2 において精度の高い IMDB，AG News 間で比較すると大きな差が存在するため，本来のタスクとしての難易度が無関係であることも確認できる．

一方で，多くのタスクにおいて Integrated Gradients，Input X Gradient が優秀であり，Saliency や Attention Weights は奮わない結果であることが確認できる．ゆえに，データセットを通して，解釈手法間の順位的な傾向は保たれていると考えることができる．

5　弱学習器による検証方法

S.Wiegreffe らは弱学習器を使用した Attention Weights の正当性検証を行っている [5]．図 4 にモデルの概要を示す．特徴マップを出力への重み付けとして使用し，LSTM 層を多層パーセプトロンに置換する．特徴マップが重要度として適切な値になっていなければ精度が大きく下がると考えられるため，一種の信頼性の評価手法として用いる．

本来は Attention Weights のための手法であるが，本研究では一般的な特徴マップを生成する手法に拡張する．そのためにいくつかの実験条件を追加する．

- 3.2 節のモデルから特徴マップを生成する
- モデルが予測したラベルに対して特徴マップを生成する (教師ラベルを使用しない)
- 特徴マップのノルムを 1.0 に正規化する

表 4　訓練済み MLP モデルにおける Test set F1 score

	IMDB	SST	AG News	20 News
Saliency	.843	.669	0.926	.723
Integrated Grad	**.887**	.725	**0.940**	.727
Input X Grad	.887	.667	0.938	**.741**
Attention	.852	.592	0.922	.660
Uniform	.862	**.728**	0.932	.524
Random	.451	.572	0.490	.453

- ベースライン手法として **Uniform**，**Random** を追加する
  - **Uniform**: 特徴マップを均一として扱う
  - **Random**: 特徴マップを毎回ランダムに生成する

5.1　実験

3 節で述べた実験設定において弱学習器を用いた検証を行う．それぞれ訓練した弱学習器において，テストセットにおける F1 score を求めたものを 表 4 に示す．

実験結果より，データセットによって傾向が異なることが確認できる．IMDB・AG News では解釈手法間の差が微小な一方で，SST・20 News では大きな差が発生することが確認できた．手法間の順位関係もデータセットによって大きく異なり，IMDB・AG News では Integrated Gradients，Input X Gradient が高い精度となるが，SST では Uniform が最も優秀になる一方で Input X Gradient の精度が低下してしまう．よって，この手法はデータセットの特性に大きな影響を受けると考えられる．また，精度の変化も微小なものであることが多い．

一方で図 5 に示すように，IMDB における訓練時の Validation Loss の推移を観測すると手法間で収束の特性が大きく異なることが確認できる．Integrated Gradients・Input X Gradient が早期に収束する一方で，Saliency・Attention Weights は収束に時間が必要となる．

収束が早い要因として，重み付けの前後の線形層を十分に訓練しなくても精度が出せることが考えられる．現に，線形層を十分に訓練する必要のある Uniform は最も収束が遅い結果となる．仮説止まりではあるものの，特徴マップがモデルの分類に有効な特徴量であるほど，収束が早くなる可能性が存在する．そのため，Loss の収束特性が意味を持つ可能性が十分に考えられる．

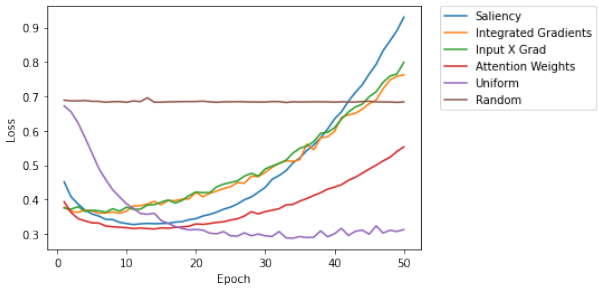

図 5　IMDB における訓練時の Validation Loss の推移

6　シャッフルを用いた検証方法

文献 [4] では Attention Weights の頑健性評価のため，シャッフルを用いた実験が行われた．Attention Weights が分類において有効な特徴量であれば，シャッフルによって出力が大きく乱れるため，モデルに対する特徴マップの信頼性を測ることが可能と考えられる．

本来は Attention Weights のための手法であるが，本研究では一般的な特徴マップを生成する手法に拡張する．実験は以下の手順で行う．

1. 弱学習器モデルを用いて入力 $\boldsymbol{x}$・特徴マップから出力ラベルの予測確率 $\boldsymbol{y}$ を求める
2. 重み付けに使用した特徴マップをランダムに入れ替える
3. 再度，出力ラベルの予測確率 $\hat{\boldsymbol{y}}$ を求める
4. 出力ラベルの変化を **Total Validation Distance** (TVD) を用いて計算する
5. 2〜4 を 100 回試行し，TVD の中央値を取る

TVD は以下の式で計算する．TVD は確率分布の距離尺度となるため，TVD が大きな値を取るほどシャッフルによって出力が乱れているとみなすことができる．

$$TVD(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{2} \sum_i |y_i - \hat{y}_i|$$

6.1　実験

それぞれの解釈手法・データセットについて実験を行い，TVD の平均値を計算した結果を表 5 に示す．

平均値の値の傾向は ERASER 法より安定するものの，同一の評価手法でもデータセットによる値の変化はある程度確認できる．例として，SST・AG News における Saliency の TVD は.093 と非常に低い値となるが，IMDB では.273 と大きい値となる．ゆえに，データセットを跨いだ値の比較検証は不可能であると言える．

表 5　データセット・解釈手法毎の TVD の平均値

	IMDB	SST	AG News	20 News
Saliency	.273	.093	.093	.164
Integrated Grad	.369	**.314**	.365	.336
Input X Grad	**.429**	.240	**.376**	**.388**
Attention	.258	.167	.161	.307

各データセットにおける順位的な関係に注目すると，Input X Gradient が特に優秀であり，次点で Integrated Gradients が高いスコアとなる．この 2 手法が優秀である点は ERASER 法や弱学習器の収束特性で見られた傾向と一致する．

7　議論

多くの評価手法・データセットに共通して Integrated Gradients と Input X Gradient は高いスコアを記録しており，手法間の性能差は小さいものになる．また，Saliency は殆どの実験で他手法より劣る結果となる．ゆえに今回の実験設定では，特徴マップの生成手法ごとに一貫した順位的な性質が確認できる．

一方で，各評価手法が同一の性質を測定しているとは断言できない．ERASER 法 / IMDB における Saliency・Attention Weights と Integrated Gradients・Input X Gradient には確たる差が存在するものの，シャッフル法 / IMDB では差が小さい．ゆえに，現状の研究成果では評価手法間に何らかの関連性が推測できるものの，同一の性質を測定していると断定することができない．

同一性を検証するためには，各評価手法がどのような性質を分析しているか調査する必要がある．今後の研究では，評価手法を最大化するような特徴マップの性質を特定するなど，より定性的な検証を中心に行う予定である．

また，データセットよりも粒度の細かい単位での検証や，モデルのパラメータの考慮など，より多くの条件で実験できるよう改良する予定である．

参考文献

[1] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Good-

fellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, Vol. 31, pp. 9505–9515. Curran Associates, Inc., 2018.

[3] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics.

[4] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics.

[6] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics.

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, Vol. abs/1312.6034, , 2013.

[8] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

[11] Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, Vol. abs/1612.08220, , 2016.

[12] Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4206–4216, Online, July 2020. Association for Computational Linguistics.