解釈可能な敵対的摂動を用いた頑健な注意機構の学習

北田 俊輔† 彌冨 仁†

† 法政大学大学院 理工学研究科 応用情報工学専攻
{shunsuke.kitada.8y@stu., iyatomi@}hosei.ac.jp

概 要

注意機構や損失勾配は入力に対する予測の説明に使われてきたが，これらの手法は摂動に頑健ではないと示唆されている．またこの 2 つの関係性の評価はこれまで順位相関が用いられてきたが，その意義にも議論の余地がある．このように，摂動に頑健な注意機構の学習方法や，注意機構と損失勾配の関係を適切に評価する方法について課題が残されている．本研究では摂動に頑健な注意機構の学習のために，interpretable adversarial training (iAdvT) をもとにした Attention iAdvT の提案を行うとともに，これらの説明手法の評価基準としてピアソン相関を用いることを主張する．4 つのオープンデータセットからなる，様々なテキスト分類タスクを用いた評価実験において，Attention iAdvT がほぼすべてのタスクで最高性能を達成した．また，注意機構と損失勾配は高く相関することを示し，すべてのタスクにおいて提案手法が一番高い相関を示すことを確認した．

1 はじめに

解釈可能なテキスト分類は現在，自然言語処理における重要な研究の 1 つである．入力文に対して予測に寄与する重要な部分を可視化または解釈するために，基本的には注意機構を基にした解釈と損失勾配を基にした解釈の 2 つの方法が用いられている．

注意機構 (Attention) [1] は深層ニューラルネットワーク (DNN) を通じて自然言語処理の分野に幅広く適用されており，多くのタスクでパフォーマンス向上に大きく貢献している．また注意機構の可視化を通じて，DNN モデルの予測時に寄与している部分を基にした説明が可能であると考えられている [2]．

予測の解釈を行う文脈においては，損失勾配に基づく方法が提案されている [3]．こうした手法は学習済みモデルの勾配に基づいて，単語列などの入力における重要性を可視化することが可能である．

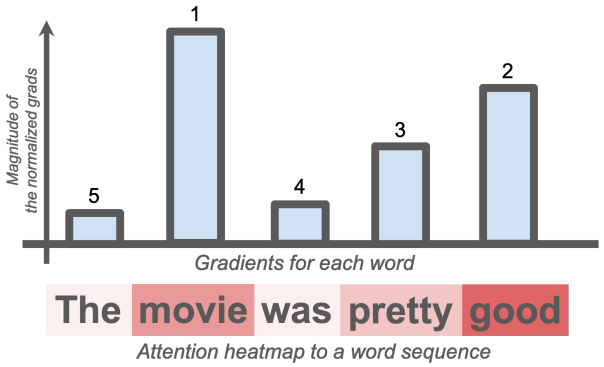


図 1: 単語列における各単語に対する勾配の大きさの棒グラフと，注意機構を可視化した際のヒートマップの例．棒グラフの上部に示す数字は各単語に対する正規化された勾配の大きさの順位を示している．一般的に，勾配を基にした可視化の順位と注意機構を基にした可視化の順位には一貫性は見られない．

注意機構に基づく解釈手法と損失勾配に基づく解釈手法の両者はそれぞれ予測の説明性を提供すると考えられているが，Jain ら [4] は，注意機構と損失勾配の間の相関係数においてしばしば相関が低いことを報告している．このとき相関係数にはケンドールの順位相関係数が用いられているが，これは順位を基にした相関係数であり，この評価指標が注意機構と損失勾配の関係を評価するのに適切かどうかは議論の余地がある．一般的に図 1 のように，勾配を元にした可視化の順位と注意機構を元にした可視化による順位には一貫性は見られない．加えて，“movie” という単語が “good” という単語よりも高い順位である必要はモデルの解釈においてあまり関係はないと考えられる．

DNN は多くの場合ノイズや摂動の影響を受けやすく，予測精度が低下する問題点がある [5]．さらにノイズや摂動を注意機構に追加すると，モデルが誤った予測をする可能性がある [4]．このような摂動は，注意機構および勾配を元にした手法による解釈を異なるものにする可能性がある．こうした摂動に対してモ

デルの頑健性を向上させるために，adversarial training (AdvT) [6] と呼ばれる正則化手法が提案されている．また，この AdvT を NLP タスクに導入する Word AdvT では，単語ベクトルに対して摂動を入力し，誤分類に対する頑健性を向上させている [7]．さらに，より解釈しやすい摂動を導入する interpretable adversarial training (iAdvT) を単語ベクトルに適用した Word iAdvT が提案され，成果を上げている [8]．

私たちは頑健な注意機構の学習を達成する普遍的な手法の提案として，単語ベクトルに対してではなく，注意機構に iAdvT を適用する Attention iAdvT を提案する．Attention iAdvT は対象タスクの分類精度の向上させると同時に，摂動に対する注意機構をより頑健にさせる．また前述の通り，私たちは注意機構と損失勾配のに対する順位相関に評価に疑問を持ったため，ピアソンの相関係数を使用してこれらの関係を再評価する意義を議論し提案する．実験セクションに提案手法および比較手法におけるピアソンの相関係数による有効性を示す．本論文での貢献は以下の通りである．

- 私たちは摂動に対する頑健な注意機構を学習する，Attention iAdvT を提案する．
- 私たちは仮説と観測に基づいた，注意機構と損失勾配の関係を評価する評価指標の再考を行う．
- 私たちは提案する Attention iAdvT を適用したモデルに対してさまざまな摂動に対して頑健になりうることを示す．

2 敵対的摂動

モデルに入力される単語列 X は T 個の単語から構成され，語彙を $\mathcal{V}$ としたときに $x_t \in \mathcal{V}$ を t 番目の単語として $X = (x_1, \cdots, x_T) = (x_t)_{t=1}^T$ のように表すことができる．また $\mathcal{Y}$ は出力のクラス数とする．

次に x_t に対応する D 次元の単語ベクトル $\boldsymbol{w} \in \mathbb{R}^D$ を考える．単語列 X に対応する単語ベクトル列 $\tilde{X}$ は $\tilde{X} = (\boldsymbol{w}_t)_{t=1}^T$ と書ける．このとき，$\tilde{Y}$ は $\mathcal{Y}$ におけるクラスラベルに対応する．したがって，N 個の訓練データ $\mathcal{D}$ は $\tilde{X}$ と $\tilde{Y}$ のペアからなる集合 $\mathcal{D} = \{(\tilde{X}_n, \tilde{Y}_n)\}_{n=1}^N$ として表すことができる．

ベースライン　Jain ら [4] にしたがって，注意機構を持つ LSTM ベースのエンコーダー **Enc** を用いて入力 $\tilde{X}$ をエンコードした．エンコーダーは隠れ層 $\boldsymbol{h}_t$ を $\boldsymbol{h}_t = \mathbf{Enc}(\boldsymbol{w}_t, \boldsymbol{h}_{t-1})$ のように各ステップ t に対して計算する．ここで $\boldsymbol{h}_0$ はゼロベクトルである．

注意機構 ϕ を考慮するため，ϕ は $\boldsymbol{h}$ および $\boldsymbol{Q} \in \mathbb{R}^D$ をスカラ値に変換する．このとき入力に対する注意は以下のように計算される:

$$\tilde{\boldsymbol{\alpha}} = \mathrm{softmax}(\phi(\boldsymbol{h}, \boldsymbol{Q})) \in \mathbb{R}^T. \tag{1}$$

ここで ϕ は additive attention [1] を考慮した．

最終的に，パラメータ $\mathcal{W}$ を持つ全結合層は予測 $\boldsymbol{q}$ を $\boldsymbol{q} = \sigma(\mathcal{W} \cdot h_\alpha) \in \mathbb{R}^{\mathcal{Y}}$ を計算する．このとき，$h_\alpha = \sum_{t=1}^T \tilde{\alpha}_t \cdot h_t$ であり，σ は活性化関数である．従って，モデルは入力 $\tilde{X}$ に対する条件付き確率 $\tilde{Y}$ を計算する:

$$p(\tilde{Y}|\tilde{X}, \mathcal{W}) = \frac{\exp(q_{\tilde{Y}})}{\sum_{m=1}^{|\mathcal{Y}|} \exp(q_m)}. \tag{2}$$

ここで q_m は $\boldsymbol{q}$ における m 番目の要素である．

テキスト分類モデルを訓練する上で，以下の最適化問題を最小化するようなモデルのパラメータを探す．

$$\hat{\mathcal{W}} = \underset{\mathcal{W}}{\mathrm{argmin}}\{\mathcal{L}(\mathcal{D}, \mathcal{W})\} \tag{3}$$

ここで $\mathcal{L}(\mathcal{D}, \mathcal{W})$ は $\ell(\tilde{X}, \tilde{Y}, \mathcal{W}) = -\log p(\tilde{Y}|\tilde{X}, \mathcal{W})$ からなる，訓練データ $\mathcal{D}$ に対する損失関数である．

$$\mathcal{L}(\mathcal{D}, \mathcal{W}) = \frac{1}{|\mathcal{D}|} \sum_{(\tilde{X}, \tilde{Y}) \in \mathcal{D}} \ell(\tilde{X}, \tilde{Y}, \mathcal{W}) \tag{4}$$

単語に対する敵対的摂動　単語に対する adversarial training (Word AdvT) [7] は提案手法である Attention iAdvT における重要な要素である．Word AdvT は目的関数に対して追加の誤差項 $\mathcal{L}_{\mathrm{AdvT}}$ を導入する．

$$\hat{\mathcal{W}} = \underset{\mathcal{W}}{\mathrm{argmin}}\{\mathcal{L}(\mathcal{D}, \mathcal{W}) + \lambda \mathcal{L}_{\mathrm{AdvT}}(\mathcal{D}, \mathcal{W})\} \tag{5}$$

ここで λ は追加した誤差項の影響を制御するハイパーパラメータであり，$\mathcal{L}_{\mathrm{AdvT}}$ は次のように定義される．

$$\mathcal{L}_{\mathrm{AdvT}}(\mathcal{D}, \mathcal{W}) = \frac{1}{|\mathcal{D}|} \sum_{(\tilde{X}, \tilde{Y}) \in \mathcal{D}} \ell(\tilde{X}^{+\boldsymbol{r}_{\mathrm{AdvT}}}, \tilde{Y}, \mathcal{W}) \tag{6}$$

$\tilde{X}^{+\boldsymbol{r}_{\mathrm{AdvT}}} = (\boldsymbol{w}_t + \boldsymbol{r}_t^{\mathrm{AdvT}})_{t=1}^T$ は単語ベクトルに対して損失が増大する方向のノイズ $\boldsymbol{r}^{\mathrm{AdvT}}$ が加えられたベクトルであり，以下のように定義される．

$$\boldsymbol{r}_t^{\mathrm{AdvT}} = -\epsilon \frac{\boldsymbol{g}_t}{||\boldsymbol{g}||_2}, \text{where } \boldsymbol{g}_t = \nabla_{\boldsymbol{w}_t} \ell(\tilde{X}, \tilde{Y}, \mathcal{W}) \tag{7}$$

ここで $\boldsymbol{g}$ は各ステップ t における $\boldsymbol{g}_t$ をすべて結合したベクトルである．

表 1: 2 値のテキスト分類タスクにおける予測精度 (F1) とピアソンの相関係数 (Corr.) の比較

Model	SST		IMDB		20News		AGNews	
	F1 [%]	Corr.	F1 [%]	Corr.	F1 [%]	Corr.	F1 [%]	Corr.
Baseline [4]	79.77	0.852	87.85	0.788	94.44	0.891	95.52	0.822
Word AdvT [7]	79.60	0.647	89.65	0.838	95.56	0.892	**95.87**	0.813
Word iAdvT [8]	79.57	0.643	89.67	0.839	95.54	0.893	95.84	0.809
Attention AdvT (**Ours**)	79.53	0.852	89.86	0.819	95.63	0.868	95.06	0.835
Attention iAdvT (**Ours**)	**82.20**	**0.876**	**90.21**	**0.861**	**95.87**	**0.897**	95.77	**0.891**

解釈可能な単語に対する敵対的摂動　単語に対する interpretable adversarial training (Word iAdvT) [8] は Word AdvT よりも解釈性の高い摂動を単語ベクトルに対して付与することでモデルを頑健にする手法である．まず $\boldsymbol{w}_k$ は語彙 $|\mathcal{V}|$ における k 番目の単語ベクトルを表す．このとき $\boldsymbol{w}_t$ から $\boldsymbol{w}_k$ に対する単語方向ベクトル $\boldsymbol{d}_{k,t}$ は次のように定義される．

$$\boldsymbol{d}_{k,t} = \frac{\tilde{\boldsymbol{d}}_{k,t}}{||\tilde{\boldsymbol{d}}_{k,t}||}, \text{where} \quad \tilde{\boldsymbol{d}}_{k,t} = \boldsymbol{w}_k - \boldsymbol{w}_t \tag{8}$$

ここですべての t および k における $\boldsymbol{d}_{k,t}$ は常に単位ベクトル $||\boldsymbol{d}_{k,t}||_2 = 1$ となる．

次に $\boldsymbol{\alpha}_t \in \mathbb{R}^{|\mathcal{V}|}$ を考える．$\alpha_{k,t}$ は $\boldsymbol{\alpha}_t$ における k 番目の要素を表し，$\boldsymbol{\alpha}_t = (\alpha_{k,t})_{k=1}^{|\mathcal{V}|}$ である．$\boldsymbol{r}(\boldsymbol{\alpha}_t)$ は $\tilde{X}$ における t 番目の単語に対して，$\boldsymbol{\alpha}_t$ による摂動である．

$$\boldsymbol{r}(\boldsymbol{\alpha}_t) = \sum_{k=1}^{|\mathcal{V}|} \alpha_{k,t} \boldsymbol{d}_{k,t} \tag{9}$$

単語ベクトル $\boldsymbol{w}$ にこの摂動が与えられたとき，$\tilde{X}^{+\boldsymbol{r}(\boldsymbol{\alpha})} = (\boldsymbol{w}_t + \boldsymbol{r}(\boldsymbol{\alpha}_t))_{t=1}^{T}$ と定義する．このとき単語ベクトルに対して損失が増大する方向のノイズ $\boldsymbol{r}^{\mathrm{AdvT}}$ は以下のように定義される．

$$\boldsymbol{r}_t^{\mathrm{AdvT}} = -\epsilon \frac{\boldsymbol{g}_t}{||\boldsymbol{g}||_2}, \ \boldsymbol{g}_t = \nabla_{\boldsymbol{\alpha}_t} \ell(\tilde{X}^{+\boldsymbol{r}(\boldsymbol{\alpha})}, \tilde{Y}, \mathcal{W}). \tag{10}$$

Word iAdvT による摂動によって，向きと大きさを単語の方向に基づいた可視化や，人間が解釈可能なシンボルに復元可能であると考えられている．

3　提案手法

本研究では，解釈可能な adversarial training を注意機構に適用することで，より頑健な注意の訓練を目指す．まず $\boldsymbol{\alpha}_k$ は注意スコア $\boldsymbol{a} = \phi(\boldsymbol{h}, \boldsymbol{Q})$ における k 番目の注意を表す．このとき，$\boldsymbol{\alpha}_t$ から $\boldsymbol{\alpha}_k$ に対する注意方向ベクトル $\boldsymbol{u}_{k,t}$ は次のように定義される．

$$\boldsymbol{u}_{k,t} = \frac{\tilde{\boldsymbol{u}}_{k,t}}{||\tilde{\boldsymbol{u}}_{k,t}||_2}, \text{where} \quad \tilde{\boldsymbol{u}}_{k,t} = \boldsymbol{a}_k - \boldsymbol{a}_t \tag{11}$$

ここですべての t および k における $\boldsymbol{u}_{k,t}$ は常に単位ベクトル $||\boldsymbol{u}_{k,t}||_2 = 1$ となる．

次に $\boldsymbol{\beta}_t \in \mathbb{R}^T$ を考える．$\beta_{k,t}$ は $\boldsymbol{\beta}_t$ における k 番目の要素を表し，$\boldsymbol{\beta}_t = (\beta_{k,t})_{k=1}^{T}$ である．$\boldsymbol{r}(\boldsymbol{\beta}_t)$ は $\tilde{X}$ における t 番目の単語に対して，$\boldsymbol{\beta}_t$ による摂動である．

$$\boldsymbol{r}(\boldsymbol{\beta}_t) = \sum_{j=1}^{T} \beta_{j,t} \boldsymbol{u}_{j,t} \tag{12}$$

注意スコア $\boldsymbol{a}$ にこの摂動が加えられたとき，$\tilde{X}^{+\boldsymbol{r}(\boldsymbol{\beta})} = (\boldsymbol{a}_t + \boldsymbol{r}(\boldsymbol{\beta}_t))_{t=1}^{T}$ と定義する．このとき注意機構に対して損失が増大する方向のノイズ $\boldsymbol{r}^{\mathrm{AdvT}}$ は以下のように定義される．

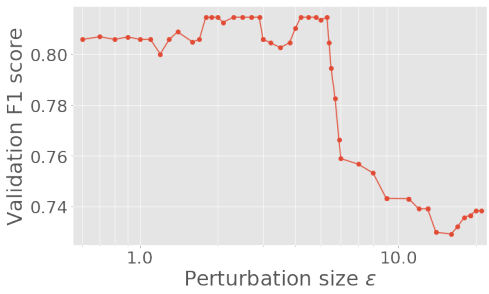
$$\boldsymbol{r}_t^{\mathrm{AdvT}} = -\epsilon \frac{\boldsymbol{g}_t}{||\boldsymbol{g}||_2}, \ \boldsymbol{g}_t = \nabla_{\boldsymbol{\beta}_t} \ell(\tilde{X}^{+\boldsymbol{r}(\boldsymbol{\beta})}, \tilde{Y}, \mathcal{W}) \tag{13}$$

4　実験

本研究で提案する Attention iAdvT を評価するため、4 つのテキスト分類タスクを用いて評価した．また、注意機構と損失勾配の関係を評価するため、ピアソンの相関係数を用いて比較した．

4.1　実験設定

Jain ら [4] と厳密に比較するため，同様のデータセットの使用および前処理の適用を行い評価に用いた．データセットは train:valid:test=6 : 2 : 2 の割合で分割を行った．提案手法である Attention iAdvT の有効性を確認するために，比較実験ではベースラインのモデルとして Jain らのモデルを使用し，敵対的摂動学習を



(a) Validation F1 scores

One of the greatest romantic comedies of the past decade

(S1-a) $\epsilon = 5$ (ハイパーパラメータ探索時の最良スコア)

One of the greatest romantic comedies of the past decade

(S1-b) $\epsilon = 16$ (ハイパーパラメータ探索時の最低スコア)

A gentle compassionate drama about grief and healing

(S2-a) $\epsilon = 5$ (ハイパーパラメータ探索時の最良スコア)

A gentle compassionate drama about grief and healing

(S2-b) $\epsilon = 16$ (ハイパーパラメータ探索時の最低スコア)

(b) 単語列に対する注意の可視化

図 2: SST [9] の valid における摂動の大小による提案手法の予測精度の変化と，ある単語列に対する注意の可視化．

用いている Word AdvT [7] および Word iAdvT [8] を比較対象として検討した．これら手法を Baseline に適用し，予測精度および注意機構と損失勾配間の相関の比較を行った．損失勾配については [3] から計算される勾配を利用した．モデルの最適化には Adam を使用し，すべての実験において $\lambda = 1$ とした．Jain ら [4] の実験では test データをパラメータ選択で用いていたが，より公正性を帰するために私たちは valid データを用いてハイパーパラメータである ϵ の探索を行った．また，同じく Jain らとは異なり，注意機構と損失勾配の関係をピアソンの相関係数で評価した．

4.2 実験結果

表 1 にテキスト分類タスクにおける予測能とピアソンの相関係数の比較結果を示す．提案手法である Attention iAdvT が予測精度および相関係数において他の手法を超える予測能を示した．ピアソンの相関係数で注意機構と損失勾配の関係を評価することで，ベースラインのモデルにおいてもこれら 2 つが強い相関を示した．特に提案手法である Attention iAdvT が他のモデルよりも強い相関を示した．

図 2 に SST [9] データセットの validation セットにおける，摂動の大きさ ϵ を変化させたときの提案手法のパフォーマンスの変化を示す．提案手法である Attention iAdvT を適用したモデルでは，摂動の大きさに関わらずほとんど一定の高い予測精度を達成し，ハイパーパラメータに対する頑健性も確認できた．また，摂動の大きさがある一定を超えると予測精度は 6% ほど急激に低下した．最良のスコアの注意の可視化と比較して，最低スコアのアテンションヒートマップはより多くのノイズを引き起こし，予測モデルに悪影響を及ぼしていることが確認できた．

5 おわりに

本研究では，解釈可能な敵対的摂動である interpretable adversarial training をもとにした Attention iAdvT を提案した．提案手法は摂動に対して頑健な注意機構の学習を可能にした．また，モデルの解釈に向けて注意機構と損失勾配がより相関するような学習を可能にした．注意機構と損失勾配の関係を適切に評価するために，先行研究で用いられていた不明瞭な評価指標の再検討を行った．この結果，注意機構と損失勾配は強く相関していることを示し，提案手法を適用することで，より強く相関することを実験で確認した．

参考文献

[1] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR preprint arXiv:1409.0473*, 2014.

[2] J. Li, W. Monroe, and D. Jurafsky, “Understanding neural networks through representation erasure,” *CoRR preprint arXiv:1612.08220*, 2016.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *Proc. of ICLR, Workshop Track*, 2013.

[4] S. Jain and B. C. Wallace, “Attention is not explanation,” in *Proc. of NAACL*, 2019, pp. 3543–3556.

[5] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere, “Did the model understand the question?” in *Proc. of ACL (Long Paper)*, ser. ACL, 2018, pp. 1896–1906.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. of ICLR, Conference Track*, 2014.

[7] T. Miyato, A. M. Dai, and I. Goodfellow, “Adversarial training methods for semi-supervised text classification,” in *Proc. of ICLR, Conference Track*, 2016.

[8] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto, “Interpretable adversarial perturbation in input embedding space for text,” in *Proc. of IJCAI*, ser. AAAI Press, 2018, pp. 4323–4330.

[9] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proc. of EMNLP*, 2013, pp. 1631–1642.