半教師あり文書分類のための仮想敵対的学習による注意機構の頑健性および解釈性の向上

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

1 はじめに

深層学習モデルは様々な分野で大きな成功を収めているにも関わらず、最先端のモデルであっても、入力される摂動に対して脆弱であることが一般的に知られている [1, 2]。近年、多くの研究が注目している注意機構 [3] においても同様の脆弱性が報告されている [4]。こうした摂動に対するモデルの頑健性向上のため、Goodfellow ら [2] は敵対的学習 (AT) を提案した。これはモデルを騙すような摂動である敵対的摂動から滑らかな識別境界の学習を期待するもので、モデルの頑健性や予測性能の向上が確認されている [5, 6]。著者らは自然言語処理 (NLP) 分野において注意機構に対する AT を提案し、複数の NLP タスクにおいて高い性能を達成するとともに、より解釈可能で明確な注意を示すことを確認した [7]。

AT はモデルの頑健性向上に寄与する一方、敵対的摂動の計算は教師あり学習への適用に限定される。後に AT を教師なしデータを利用可能な半教師あり設定へと拡張した仮想敵対的学習 (VAT) が提案され [8]、優れた成果が報告されている [6, 9]。この手法は現在の予測に基づいて敵対的摂動の方向を計算することで、教師なしデータを含むすべてのデータを活用して識別境界を滑らかにする効果があり、結果的にモデルの頑健性向上が期待できる。一方で、図 1 に示すような注意機構に対する VAT の有効性について、特に NLP タスクの観点から同様の効果を示した研究は著者らが知る限り存在しない。

本論文では、VAT に基づいた注意機構のための仮想敵対的学習である仮想敵対的注意学習 (*Attention VAT*) ならびに、より解釈可能性の高い注意機構のための敵対的学習 [7] を半教師あり学習である VAT で拡張した *Attention iVAT* を提案する。我々の Attention VAT/iVAT は、著者らの提案する仮説である、予測に寄与しうる重要な単語を見つけるためには単語の埋め込みそのものよりも注意の重みが重要である [7] という主張に基づいている。我々の学習法は注意機構を有する様々なモデルに適用可能な汎用性の高い学習法である。

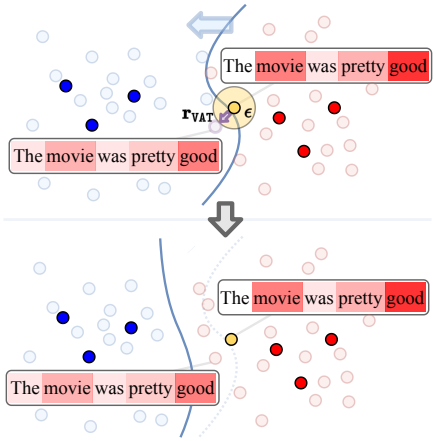


図 1: 提案法である仮想敵対的注意学習 (Attention VAT/iVAT) の概念図。本手法は利用できるすべてのデータから仮想的な敵対的摂動 r_{VAT} を計算することで識別境界の平滑化を期待する。このときの摂動は注意機構に対するノルム ϵ によって定義される。

我々の提案の効果を実証するために、広く評価に利用されている文書分類データセットを用いて最新の AT/VAT を元にした学習法と比較した。また、提案法で得られた注意の重みが (i) 勾配によって計算される単語の重要度と、(ii) 人手によってアノテーションされた予測根拠との一致度を評価した。

本研究の貢献　我々の注意機構に対する仮想敵対的学習を評価した結果、次のような知見が得られた: **(1)** 従来の AT ベースの手法だけでなく、最新の VAT ベースの手法と比較して、半教師あり設定下での予測性能が有意に優れていた; **(2)** 学習された注意の重みが勾配を元にした単語の重要度とより強い相関を示し、人手による予測根拠とより良い一致を示した; **(3)** 教師なしデータの量を最大 7 倍程度まで増やすことで予測性能が向上した。

2　注意機構のための仮想敵対的学習

本章では、提案する Attention VAT および Attention iVAT と呼ぶ仮想敵対的注意学習について説明する。提案法を適用するモデルの詳細は付録 A に示す。

2.1　仮想敵対的注意学習

Attention VAT は、VAT [8] から着想を得たものであり、著者らが提案した Attention AT [7] の拡張と捉えることができる。AT が教師ありデータに基づいて敵対的摂動を決定するのに対し、VAT は教師なしデータからも “仮想的” な敵対的摂動を計算できる。このとき注意機構に対する仮想敵対的摂動は、元の学習サンプルに対するモデルの予測出力の分布と、元の学習サンプルに摂動を加えた場合に対するモデルの予測出力の分布との KL ダイバージェンスを最大化する方向の摂動であり、注意機構に対する最悪の摂動として定義される。

本研究では半教師あり学習の設定として、教師ありデータからなる集合 $\mathcal{D}$ と教師ありおよび教師なしデータからなる集合 $\mathcal{D}'$ を学習対象とする。なお教師ありデータ数を N_{l} と教師なしデータ数を N_{ul} と表す。摂動 $\boldsymbol{r}$ が注意スコアに加えられた際の入力単語列 $X_{\tilde{\boldsymbol{a}}}$ を、$X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}$ として表す。Attention VAT は注意スコア付き入力単語列に対して仮想敵対的摂動 $\boldsymbol{r}_{\mathrm{VAT}}$ を推定するために、以下の損失項を最小化する:

$$\mathcal{L}_{\mathrm{VAT}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathrm{VAT}}}; \hat{\boldsymbol{\theta}}) = \frac{1}{|\mathcal{D}'|} \sum_{X \in \mathcal{D}'} \mathcal{L}_{\mathrm{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathrm{VAT}}}; \hat{\boldsymbol{\theta}}), \tag{1}$$

$$\boldsymbol{r}_{\mathrm{VAT}} = \underset{\boldsymbol{r}:||\boldsymbol{r}||_2 \le \epsilon}{\operatorname{argmax}} \mathcal{L}_{\mathrm{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}; \hat{\boldsymbol{\theta}}), \tag{2}$$

$|\mathcal{D}'| = N_{\mathrm{l}} + N_{\mathrm{ul}}$ は学習対象の全データ数、ϵ は摂動のノルムを制御するハイパーパラメータ、$\hat{\boldsymbol{\theta}}$ は現在のモデルの全パラメータである。ここで $\mathcal{L}_{\mathrm{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}; \hat{\boldsymbol{\theta}})$ は $\mathrm{KL}(\cdot||\cdot)$ で計算される KL ダイバージェンスである:

$$\mathcal{L}_{\mathrm{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathrm{VAT}}}; \hat{\boldsymbol{\theta}}) = \mathrm{KL}(p(\cdot|X_{\tilde{\boldsymbol{a}}}, \hat{\boldsymbol{\theta}})||p(\cdot|X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}_{\mathrm{VAT}}}, \hat{\boldsymbol{\theta}})). \tag{3}$$

学習時には、入力データの注意スコア $\tilde{\boldsymbol{a}}$ に対する仮想敵対的摂動 $\boldsymbol{r}_{\mathrm{VAT}}$ を現在のモデルパラメータ $\hat{\boldsymbol{\theta}}$ に基づいて計算し、その注意スコアへ付加する:

$$\tilde{\boldsymbol{a}}_{\mathrm{vadv}} = \tilde{\boldsymbol{a}} + \boldsymbol{r}_{\mathrm{VAT}}. \tag{4}$$

2.2　解釈可能な仮想敵対的注意学習

Attention iVAT は解釈可能な単語埋め込みに対する敵対的学習 (Word iVAT) [6] から着想を得ており、著者ら [7] によって提案された Attention iAT に対する半教師あり学習への拡張とみなすことができる。Attention iAT と同様に我々の Attention iVAT は各単語への注意の違いを利用する一方で、我々の手法は多くの教師なしデータを扱うことで予測性能とモデルの解釈性が更に向上することを期待している。

まず $\boldsymbol{d}_t \in \mathbb{R}^T$ を、文中の t 番目の単語に対する注意スコア $\tilde{a}_t$ と任意の k 番目の単語に対する注意スコア $\tilde{a}_k$ との単語の注意差ベクトルを定義する:

$$\boldsymbol{d}_t = (d_{t,k})_{k=1}^{T} = (\tilde{a}_t - \tilde{a}_k)_{k=1}^{T}. \tag{5}$$

このベクトルを用いて、t 番目の単語に対する正規化された単語注意差ベクトルを計算する:

$$\tilde{\boldsymbol{d}}_t = \boldsymbol{d}_t / ||\boldsymbol{d}_t||_2. \tag{6}$$

t 番目の単語に対する注意の摂動 $r(\boldsymbol{\alpha}_t)$ を、訓練可能なパラメータ $\boldsymbol{\alpha}_t = (\alpha_{t,k})_{t=1}^{T} \in \mathbb{R}^T$ と正規化された注意度差ベクトル $\tilde{\boldsymbol{d}}_t$ を使って $r(\boldsymbol{\alpha}_t) = \boldsymbol{\alpha}_t^{\top} \cdot \tilde{\boldsymbol{d}}_t$ と定義する。全ての t に対する $\boldsymbol{\alpha}_t$ を $\boldsymbol{\alpha}$ として、入力文に対する敵対的摂動 $\boldsymbol{r}(\boldsymbol{\alpha})$ を計算する:

$$\boldsymbol{r}(\boldsymbol{\alpha}) = (r(\boldsymbol{\alpha}_t))_{t=1}^{T}. \tag{7}$$

式 2 の $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}}$ と同様な $X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{\alpha})}$ を導入し、損失が最大になる最悪の注意差ベクトルの方向を計算する:

$$\boldsymbol{r}_{\mathrm{iVAT}} = \underset{\boldsymbol{r}:||\boldsymbol{r}||_2 \le \epsilon}{\operatorname{argmax}} \mathcal{L}_{\mathrm{KL}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}+\boldsymbol{r}(\boldsymbol{\alpha})}; \hat{\boldsymbol{\theta}}). \tag{8}$$

式 4 と同様に、注意スコア $\tilde{\boldsymbol{a}}$ に対して計算した敵対的摂動 $\boldsymbol{r}_{\mathrm{iVAT}}$ を付加する:

$$\tilde{\boldsymbol{a}}_{\mathrm{ivadv}} = \tilde{\boldsymbol{a}} + \boldsymbol{r}_{\mathrm{iVAT}}. \tag{9}$$

2.3　仮想敵対的学習によるモデルの学習

学習時には、現在のモデル $\hat{\boldsymbol{\theta}}$ に基づいて、仮想敵対的摂動を生成する。このとき生成した摂動を元に、以下の 2 つの損失項を最小化する:

$$\tilde{\mathcal{L}} = \underbrace{\mathcal{L}(X_{\tilde{\boldsymbol{a}}}, \boldsymbol{y}; \hat{\boldsymbol{\theta}})}_{\substack{\text{The loss from} \\ \text{unmodified examples}}} + \underbrace{\lambda \mathcal{L}_{\mathrm{VAT}}(X_{\tilde{\boldsymbol{a}}}, X_{\tilde{\boldsymbol{a}}_{\mathrm{VADV}}}; \hat{\boldsymbol{\theta}})}_{\substack{\text{The loss from} \\ \text{virtual adversarial examples}}}, \tag{10}$$

λ は各損失項を制御するハイパーパラメータである。ここで、$X_{\tilde{\boldsymbol{a}}_{\mathrm{VADV}}}$ は Attention VAT において $X_{\tilde{\boldsymbol{a}}_{\mathrm{vadv}}}$ であり、Attention iVAT において $X_{\tilde{\boldsymbol{a}}_{\mathrm{ivadv}}}$ である。

3 実験設定

本章では、評価に使用するモデルの設定やデータセット、評価基準について述べる。

3.1 モデルの設定と評価データセット

付録 A に示すベースラインモデルに対して、我々の提案する 2 つの手法 Attention VAT/iVAT と最新の AT/VAT 手法を教師ありおよび半教師あり設定で比較した。比較手法の詳細を付録 B に示す。評価用データセットとして、教師ありデータセット [10, 11]、教師なしデータセットとして [12] を使用し、学習用、検証用、テスト用にそれぞれ分割した。これらの詳細を付録 C に示す。

3.2 評価基準

予測性能　提案法および比較法によって学習させたモデルの予測性能を比較した。評価指標として、[4, 7] に従って F1 スコアを用いた。各モデルはそれぞれ教師ありおよび半教師ありの設定で評価した。

単語の重要度との相関　提案法および比較法を適用したモデルによって得られる注意の重みと勾配を元にした単語の重要度 [13] との一致度を比較した。この一致度を評価するために、注意の重みと単語の重要度との間のピアソン相関で比較した。

予測根拠の再現性　提案法および比較法を適用したモデルによって得られる注意の重みと、人手によってアノテーションされた予測根拠の一致度を ERASER ベンチマーク [14] を用いて比較した。

教師なしデータの効果　半教師あり学習における教師なしデータの効果を理解するため、教師なしデータの量とモデルの予測性能の関係を分析した。

4 実験結果

本章では、前章に示した評価基準を元に実験した結果を共有する。表 1 は教師ありおよび半教師あり設定における各モデルの (1) 予測性能 (F1 スコア)、および (2) 注意の重みと勾配を元にした単語の重要度との間のピアソン相関 (Corr.) を示す。

予測性能　教師ありの設定において、提案法である Attention VAT/iVAT は (1) vanilla モデルと比較して明らかな優位性を示し、(2) 単語埋め込みに対する AT (Word AT/iAT) [5, 6] と比較しても優れた性能を示した。さらに (3) 注意機構に対する AT (Attention AT/iAT) [7] と同等の性能を示した。以上より、注意機構に対する AT/VAT の有効性を確認できた。

表 1: 教師ありおよび半教師ありモデルに対する、予測性能 (F1 スコア) および注意と勾配を元にした単語の重要度のピアソン相関係数 (Corr.) の比較

Model	SST		IMDB	
	F1 [%]	Corr.	F1 [%]	Corr.
Vanilla [4]	79.27	0.852	88.77	0.788
Word AT [5]	79.61	0.647	89.65	0.838
Word iAT [6]	79.57	0.643	89.64	0.839
Word VAT [8]	81.90	0.651	89.79	0.841
Word iVAT [6]	81.98	0.648	89.78	0.843
Attention AT [7]	81.72	0.852	90.00	0.819
Attention iAT [7]	82.20	0.876	**90.21**	**0.861**
Attention VAT (**Ours**)	82.19	0.878	90.20	0.859
Attention iVAT (**Ours**)	**82.21**	**0.881**	90.18	0.857

(a) 教師ありモデル

Model	SST $N_{\mathrm{ul}} = 50{,}000$		IMDB $N_{\mathrm{ul}} = 150,000$	
	F1 [%]	Corr.	F1 [%]	Corr.
Word VAT [8]	83.04	0.779	92.21	0.853
Word iVAT [6]	83.07	0.781	92.30	0.859
Attention VAT (**Ours**)	83.18	0.898	92.48	0.879
Attention iVAT (**Ours**)	**83.22**	**0.901**	**92.56**	**0.883**

(b) 半教師ありモデル

半教師ありの設定において、Word VAT/iVAT は Word AT/iAT より、また提案する Attention VAT/iVAT も同様に Attention AT/iAT より有意な予測性能の向上ならびに勾配に基づく単語の重要度とより高い相関を示し、特に提案法は最もよい結果となった。各タスクの学習に追加で教師なしデータを利用することで、我々の手法を適用したモデル、特に Attention iVAT では予測性能が大幅に向上したと考えられる。

単語の重要度との相関　注意の重みと単語の重要度の相関については、我々の Attention VAT/iVAT で得られた単語への注意は、勾配によって計算された単語の重要度と強く相関していることを確認した。この傾向は我々の手法のヒントとなった [7] の Attention AT/iAT においても報告されている。我々は、VAT を用いた半教師あり設定がこれらの相関も顕著な影響を与えていることを観測した。こうした相関は我々の提案法により強化されたと言える。

予測根拠の再現性　表 2 は vanilla モデルと半教師ありモデルにおける根拠選択の予測性能を示す。提

表 2: 根拠選択における各モデルの予測性能

Model	AUPRC	Avg. Prec.	ROC-AUC
Vanilla [4]	0.326	0.395	0.563
Word VAT [5]	0.349	0.413	0.581
Word iVAT [6]	0.350	0.414	0.582
Attention VAT (**Ours**)	0.403	0.477	0.646
Attention iVAT (**Ours**)	**0.417**	**0.489**	**0.651**
†DeYoung ら [14]	0.502	-	-

† DeYoung ら [14] は根拠選択に最適化されたモデルを構築しているのに対し、我々の手法を適用したモデルは根拠選択を直接最適化しているわけではない。

案する Attention VAT/iVAT は注意の重みを元にした根拠選択において、人手のアノテーションとの一致度が高いことを示した。他の AT/VAT ベースの手法と比較して、特に我々の Attention iVAT が全ての評価基準において優れていることが分かった。

教師なしデータの効果　図 2 は教師なしデータの量とモデルの検証スコアの関係を示す。VAT ベースの手法において、教師なしデータの数が教師ありデータの数の約 7 倍程度になるまで予測性能の向上を確認した。本実験において我々が教師なしデータとして使用したソースは、元のデータセットとは性質や品質が異なるものである (詳細は付録 C を参照)。それにも関わらず提案法がモデルの性能と解釈性の両者を向上させたことは注目に値する。

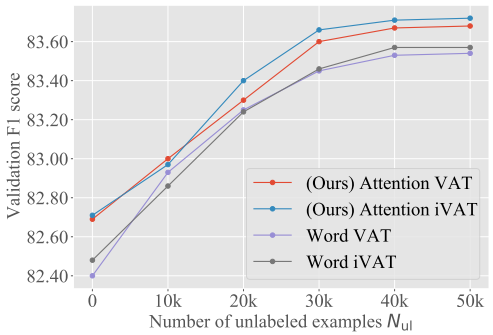
5　議論と今後の展望

本研究では、Attention VAT および Attention iVAT と呼ぶ仮想敵対的注意学習を提案し、教師ありおよび半教師あり設定にて提案法の有効性を確認した。

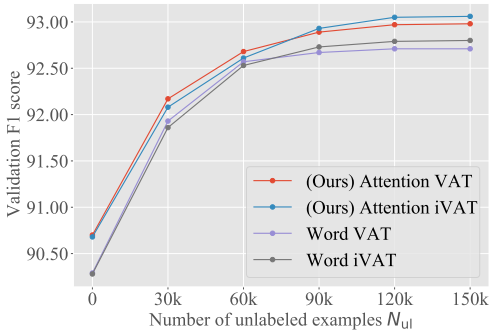
5.1　注意機構に対する AT/VAT

教師あり設定において、我々の VAT ベースの手法は 2 つの側面 (1) 予測性能 (2) 単語の重要度との相関で、AT ベースの手法と同程度向上した。AT は実際の教師情報を使用する一方で、VAT はモデルの予測出力を利用している。そのため教師あり設定において AT は VAT よりも良い性能であることが期待されるが、VAT が同程度の性能を表したことにより、提案法に対して否定的な結果は得られなかった。

半教師あり設定において、我々の VAT ベースの手法は上記 2 つの側面を更に効果的に向上させていることを確認した。これは多くの教師なしデータを利用することで、識別境界をより滑らかにする VAT の効果を更に高めていると考えられる。特に単語の



(a) SST (N_l = 6,920)



(b) IMDB (N_l = 17,186)

図 2: SST [10] と IMDB [11] に対する教師なしデータ [12] の量と各学習法による検証スコアの関係

埋め込みではなく注意機構に VAT を適用することで、モデルの頑健性 (予測性能) が向上し、さらには解釈性 (人手による根拠との一致度) も向上した。

実験では、提案法を含む VAT ベースの手法において、教師ありデータと異なるソースを教師なしデータとして使用したにもかかわらず好ましい性能を示した。これは提案法の汎用性の高さを示していると考えられる。今後は、教師ありデータに教師なしデータを構築し、提案法の効果を検証したい。

5.2　人手による根拠との一致性

我々の提案法である Attention iVAT は、他の手法と比較して人手による根拠のアノテーションとの一致率が高かった。これは式 6 に示すように、Attention iVAT では単語注意差のノルムは 1 に正規化されるため、各単語に対する注意度の差異がより明確になるとともに、より効果のある仮想的な敵対的摂動が得られているためと考えられる。この類似の性質は著者ら [7] の Attention iAT においても同様の議論がされている。さらにこうした明確な注意と仮想敵対的摂動に伴う注意は、言語の解析や理解に必要な注意の普遍的な特徴を学習すると期待できるため、今後さらなる分析を行っていきたい。

参考文献

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR, Conference Track Proceedings*, 2013.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR, Conference Track Proceedings*, 2014.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[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)*, Association for Computational Linguistics (ACL), pp. 3543–3556, 2019.

[5] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR, Conference Track Proceedings*, 2016.

[6] Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, AAAI Press, pp. 4323–4330, 2018.

[7] Shunsuke Kitada and Hitoshi Iyatomi. Attention meets perturbations: Robust and interpretable attention with adversarial training. *CoRR arXiv:2009.12064*, 2020.

[8] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, Vol. 41, No. 8, pp. 1979–1993, 2018.

[9] Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. Seqvat: Virtual adversarial training for semi-supervised sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), pp. 8801–8811, 2020.

[10] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics (ACL), pp. 1631–1642, 2013.

[11] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 of *Association for Computational Linguistics (ACL)*, pp. 142–150, 2011.

[12] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, International Speech Communication Association (ISCA), pp. 2635–2639, 2014.

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR, Workshop Track Proceedings*, 2013.

[14] 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*, Association for Computational Linguistics (ACL), pp. 4443–4458, 2019.

A　Vanilla モデルの構築と学習

本章では、ベースラインとなる vanilla モデルについて述べる。このモデルは先行研究 [4, 7] がその提案の有効性を示すために使われている。本研究では広範囲に渡って検証がされているその recurrent neural network (RNN) を元にしたモデルをベースラインとして使用した。

A.1　注意機構を有するモデルの構築

まず X を one-hot エンコーディングされた単語列 $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) \in \mathbb{R}^{|V| \times T}$ と表す。なお T は単語列中の単語数、$|V|$ は語彙数を表す。ここで、単語列を表す表現の省略形である $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T)$ as $(\boldsymbol{x}_t)_{t=1}^T$ を導入する。次に、$\boldsymbol{x}_t$ に対応する d 次元の単語埋め込み表現を $\boldsymbol{w}_t \in \mathbb{R}^d$ とする。このとき、単語列に対する単語埋め込みは $(\boldsymbol{w}_t)_{t=1}^T \in \mathbb{R}^{d \times T}$ と表す。この単語埋め込み表現を双方向 RNN **Enc** を用いて m 次元の隠れ表現 $\boldsymbol{h}_t$ としてエンコードする: $\boldsymbol{h}_t = \mathbf{Enc}(\boldsymbol{w}_t, \boldsymbol{h}_{t-1})$。次に、[3] が提案した additive function を用いて、t 番目の単語に対する注意スコア $\tilde{a}_t$ を $\tilde{a}_t = \boldsymbol{c}^\top \tanh(W\boldsymbol{h}_t + \boldsymbol{b})$ として定義する。なお $W \in \mathbb{R}^{d' \times m}$ と $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^{d'}$ はモデルパラメータである。そして、文全体に対する注意スコア $\tilde{\boldsymbol{a}} = (\tilde{a}_t)_{t=1}^T$ から、全単語の注意の重み $\boldsymbol{a} \in \mathbb{R}^T$ を計算する。

$$\boldsymbol{a} = (a_t)_{t=1}^T = \mathrm{softmax}(\tilde{\boldsymbol{a}}). \tag{11}$$

注意の重み $\boldsymbol{a}$ と隠れ表現 $\boldsymbol{h}_t$ を用いて加重平均 $\boldsymbol{h}_{\boldsymbol{a}}$ を計算する: $\boldsymbol{h}_{\boldsymbol{a}} = \sum_{t=1}^T a_t \boldsymbol{h}_t$。上記の $\boldsymbol{h}_{\boldsymbol{a}}$ は全結合層 Dec へと入力され、最終的な予測値を出力する: $\hat{\boldsymbol{y}} = \sigma(\mathbf{Dec}(\boldsymbol{h}_{\boldsymbol{a}})) \in \mathbb{R}^{|\boldsymbol{y}|}$。このとき σ は活性化関数であり、$|\boldsymbol{y}|$ は予測するクラスの数である。

A.2　注意機構を有するモデルの学習

学習時には、モデルは注意スコア $\tilde{\boldsymbol{a}}$ である単語列 X である $X_{\tilde{\boldsymbol{a}}}$ から、$\boldsymbol{y}$ の条件付き確率を学習する: $p(\boldsymbol{y}|X_{\tilde{\boldsymbol{a}}}; \boldsymbol{\theta})$。ここで、$\boldsymbol{\theta}$ はモデルの全パラメータである。このとき、(教師ありの) 学習データ $\mathcal{D}$ を用いて損失項である以下の負の対数尤度を最小化する:

$$\mathcal{L}(X_{\tilde{\boldsymbol{a}}}, \boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{(X, \boldsymbol{y}) \in \mathcal{D}} -\log p(\boldsymbol{y}|X_{\tilde{\boldsymbol{a}}}; \boldsymbol{\theta}). \tag{12}$$

B　モデルの設定

我々は提案法である Attention VAT/iVAT を含む、最新の AT および VAT ベースの学習法を比較した。これらは付録 A に示すベースラインに対して適用しており、平等な比較のために [4, 7] の実験設定に則った。本研究では、教師ありおよび半教師あり設定の下、以下のモデルを比較した:

教師ありモデル　vanilla [4]、Word AT [5]、Word iAT [6]、Attention AT [7]、Attention iAT [7]、そして提案法である Attention VAT および Attention iVAT

半教師ありモデル　として、Word AT [5]、Word iVAT [6]、そして提案法である Attention VAT および Attention iVAT

表 3: 評価用データセットの統計量。評価には一般的に知られているデータセットを使用した。データセットを学習 (train) セット、検証 (valid) セット、評価 (test) セットに分割した。分割方法と前処理は著者ら [7] の先行研究と同様である。

Dataset	SST [10]	IMDB [11]
クラス数	2	2
学習用データ数	6,920	17,186
検証用データ数	872	4,294
テスト用データ数	1,821	4,353
語彙数	13,723	12,485
平均 単語/文 数	18	171

C　データセットとタスク

表 3 は教師ありデータの統計量を示す。我々は [4] および [7] に従って、実験で使用するデータセットを学習用、検証用、テスト用に分割し、それぞれ前処理を実施した。本研究では、以下の教師ありおよび教師なしデータセットを使用した:

教師ありデータセット　Standard Sentiment Treebank (**SST**) [10], **IMDB**, a large movie reviews corpus [11] を使用した。これらのデータセットは文がポジティブかネガティブかの感情が付与されており、これらを予測するようにモデルを学習させた。

教師なしデータセット　One Billion Word Language Model Benchmark [12] を半教師あり学習のためのラベルなしデータとして利用した。このベンチマークは言語モデルを評価するために幅広く利用されており、近年の VAT ベースの研究である [9] においても使用されている。従って、我々はこのデータセットが私達の実験においても有効であると考え採用した。教師ありデータサイズが比較的小さいことを考慮して、教師なしデータの 1%を無作為にサンプリングし、半教師あり学習における教師なしデータとして使用した。