根拠箇所に基づく自動採点結果の説明

佐藤 汰亮 [1,2]　舟山 弘晃 [1,2]　塙 一晃 [1,2]　浅妻 佑弥 [1]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所
{tasuku.sato.p6, h.funa, asazuma.yuya.r7}@dc.tohoku.ac.jp
inui@tohoku.ac.jp
kazuaki.hanawa@riken.jp

概要

深層学習の進歩に伴い，自動採点の性能は向上した．しかし，深層学習モデルは内部挙動がブラックボックスであるため，採点の根拠を説明できない．本研究では，根拠箇所提示問題のタスクを説明性の観点から再定義し，複数の損失関数と複数の根拠箇所提示手法の組み合わせで複数パターンの実験を行い，説明性の高い根拠箇所提示手法を確かめた．実験により，自動採点の分野では注意重みは最も妥当性が高く，勾配を用いた手法と同程度に忠実性が高いことがわかった．また，Integrated Gradients は最も忠実性が高く，注意重みと同程度に妥当性が高いことがわかった．ここから，自動採点の分野では，説明性における妥当性と忠実性，どちらの観点を重要視するかを元に根拠箇所提示手法を使い分けることが重要であることがわかった．

1　はじめに

自動採点とは，ある設問に解答した答案に対し，自動で得点を付与するタスクである．図 1 に例を示す．自動採点は，教師の採点労働を軽減できることや，人手による採点ブレがなく，公正な採点を行えることから，教育現場での需要があり，研究が行われている [1, 2, 3]．また，答案に対して即時にフィードバックを返せることから，自学学習のためのオンラインツールとしての実応用化が進んでいる [4]．

深層学習の進歩に伴い，自動採点の性能は向上した [1, 2, 3]．しかし，深層学習モデルは内部挙動がブラックボックスであり，採点の根拠を説明できない．自動採点の分野では採点基準が厳密に定まっているため，モデルが採点の根拠を説明できないことはユーザの不信感につながる．また開発者としても，採点の根拠を説明できないモデルはデバッグを難解にする．以上の観点から，採点根拠の説明性は採点性能と同等に重要な要素である．

深層学習モデルの意思決定を人間が解釈できる形で提示することは，法律や医療など，様々な分野で重要視されている [5, 6]．このような性質を**説明性**と呼ぶ．説明性の高いモデルの需要は高まっており，XAI [7] の文脈で活発に研究が行われている．

Jacovi ら [8] は，説明性を**妥当性（plausibility）**と**忠実性（sufficiency）**の 2 つの観点でまとめている．妥当性とは，予測に対する説明が，人間にとってどの程度納得のいくものであるかを表す概念である．忠実性とは，予測に対する説明が，モデルの予測過程をどの程度反映しているかを表す概念である．

Mizumoto ら [2] は，自動採点における得点予測根拠の提示性能を評価するため，**根拠箇所提示問題**（2 節参照）を定義した．彼らは，注意層を根拠箇所ラベルで教師あり学習するモデルとしないモデル（3.4 節参照）を比較した結果，前者は後者と比較し，根拠箇所提示性能が高いことを示した．

しかし Mizumoto ら [2] は，根拠箇所提示手法を真の根拠箇所との一致度（妥当性）のみで評価しており，忠実性に対する評価を行っていない．また，根拠ラベルを用いた学習手法や根拠箇所提示手法が複数考えられる中，彼らは一つの手法のみを取り上げており，十分な評価が行われているとはいえない．

本研究では，Mizumoto ら [2] による根拠箇所提示問題のタスクを，説明性 [8] の観点から再定義した．また，複数の損失関数と複数の根拠箇所提示手法の組み合わせで複数パターンの実験を行い，説明性の高い根拠箇所提示手法を確かめた．実験の結果，注意重みは最も妥当性が高く，勾配を用いた手法と同程度に忠実性が高いことがわかった．また Integrated Gradients は最も忠実性が高く，注意重みと同程度に妥当性が高いことがわかった．以上のことから，自動採点においては，説明性のどの観点を重要視するのかによって根拠箇所提示手法を選択す

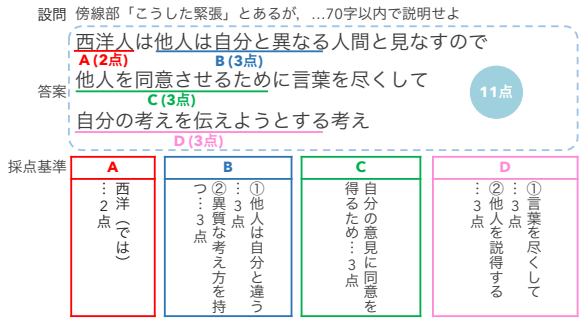

図 1　自動採点のタスク概要図．答案中の下線部は根拠箇所であり，採点基準の情報を包含した箇所を表す．

ることが重要であることがわかった．

2　問題設定

本研究では，自動採点における根拠箇所提示問題に取り組む．

まず自動採点とは，答案 $\boldsymbol{x} = (x_1, x_2, ..., x_t)$ に対し，その得点 $s \in \{0, 1, ..., S\}$ を出力するタスクである．ここで，x_t は t 番目の単語，S は配点である．また，一つの設問に対して複数の採点項目が存在し，各採点項目の配点の合計点がその答案の得点となる．採点性能は QWK [9] によって測る．

根拠箇所提示問題とは，入力答案 $\boldsymbol{x}$ において，得点予測の根拠箇所ラベル $\boldsymbol{r} = (r_1, r_2, ..., r_t)$，$r_t \in \{0, 1\}$ を提示するタスクである．ここで，$r_t = 1$ のとき，単語 x_t が根拠箇所とみなされる．モデルから任意の方法で生成された特徴マップを $\boldsymbol{f} = (f_1, f_2, ..., f_t)$，$0 < f_t < 1$，$\sum \boldsymbol{f} = 1$，その最大値を f_{max}，閾値を $h \in \{a \in R | 0 \leq a \leq 1\}$ とすると，r_t は $f_{max} - f_t < h$ のとき 1，それ以外は 0 である．

生成された根拠箇所は，妥当性と忠実性 [8] の観点から評価を行う．詳細は 3.5 節で述べる．

3　実験設定

本研究では，複数の損失関数と複数の根拠箇所提示手法の組み合わせで複数パターンの実験を行い，最も説明性の高い根拠箇所提示手法を確かめる．

3.1　データセット

本実験では，2022 年 2 月に公開を予定している，NII 国立研究所理研記述問題採点データセット第 2 版を使用する．[1)] このデータセットは，受験者が記述した答案テキストと，採点者によって付与された得点及び，その得点を付与した根拠となる根拠箇所の 3 つ組で構成されている．各設問には複数の採点項目が存在し，それぞれの項目毎にアノテーションが付与されている．今回は，採点項目ごとに独立してモデルを学習し，評価を行う．

このデータセットにおける自動採点は加点式の情報包含問題であるため，0 点答案には根拠箇所のラベルが付与されていない．そのため，0 点答案が全体の半分以上を占める採点項目は実験の対象から除外する．実験はデータを学習データ 200 件，開発データ 100 件，評価データ 200 件に分割し，33 個の採点項目を対象に行う．

1)　第 1 版のデータは以下の URL からダウンロード可能である．https://www.nii.ac.jp/dsc/idr/rdata/RIKEN-SAA/

3.2　モデル

実験には Riordan ら [1] を参考に，単語埋め込み + biLSTM + 注意層の分類器モデルを使用する．単語埋め込みは MeCab[10] で単語分割された Wikipedia のデータを用い，word2vec で学習されたものを使用する．モデルパラメータの詳細や注意層の詳細は参考情報の A.1 節を参照されたい．

3.3　特徴マップ

根拠箇所提示手法として，以下の手法によって生成された特徴マップを使用する．勾配から特徴マップを生成する手法では，各入力単語に対して得られるベクトルのノルムをとり，特徴マップとする．

注意重み　注意層における注意重みを，予測に対する各単語の重要度だとみなし，特徴マップとして使用する手法．

Saliency Map [11]　各入力の損失に対する勾配から生成する手法．

Input X Gradient [12]　Saliency Map の特徴マップに対して，入力を乗算する手法．入力の特徴量を反映した特徴マップの生成が可能となる．

Integrated Gradients [13]　勾配をベースラインから入力方向へ積分し，生成する手法．Implementation Invariance と Seisitivity の公理を満たし，説明性の高い特徴量を生成できるとされている．

Random　一様分布から特徴マップを生成する手法．

3.4　損失関数

自動採点データセットには得点ラベルと 根拠箇所ラベルが含まれるが．これらの信号を用いたモデルの学習方法として，以下の方法が考えられる．

得点学習　モデルの最終出力と得点ラベル間の交差

表 1 損失関数ごとの QWK の結果

uns	attn	grad	attn&grad
0.739	0.774	0.771	**0.782**

表 2 根拠箇所の一致度（妥当性，↑）の結果

特徴マップ \ モデル	uns	attn	grad	attn&grad
注意重み	0.477	**0.793**	0.549	0.788
Input X Gradient	0.462	0.629	0.692	0.724
Integrated Gradients	0.462	0.558	0.753	0.753
Saliency Map	0.469	0.642	0.672	0.698
Random	0.356	0.357	0.357	0.356

表 3 削除率（忠実性，↓）の結果

特徴マップ \ モデル	uns	attn	grad	attn&grad
注意重み	0.226	0.230	0.158	0.163
Input X Gradient	0.257	0.230	0.177	0.192
Integrated Gradients	**0.196**	**0.170**	**0.144**	**0.140**
Saliency Map	0.297	0.300	0.201	0.240
Random	0.590	0.599	0.530	0.558

エントロピー誤差をとる．

注意層学習　注意層の重みを根拠箇所ラベルで教師あり学習する．

勾配ノルム学習　各入力の損失に対する勾配のノルムを根拠箇所ラベルで教師あり学習する．

勾配ノルム学習を行なった先行研究は存在しないが，勾配を教師あり学習する研究は存在し，タスクの性能が向上することが報告されている [14]．よって，勾配ノルム学習も注意層学習と同じ枠組みで学習できると考えられる．勾配ノルム学習に用いる特徴マップには Saliency Map や Input X Gradient なども考えられるが，今回は最も説明性が高いとされている Integrated Gradients を使用する．各学習方法の損失関数 $\mathcal{L}_{score}$，$\mathcal{L}_{attn}$，$\mathcal{L}_{grad}$ の詳細は参考資料の A.2 節を参照されたい．

本実験では，これらの損失関数の組み合わせとして，1：$\mathcal{L}_{score}$（**uns**），2：$\mathcal{L}_{score}+\mathcal{L}_{attn}$（**attn**），3：$\mathcal{L}_{score}+\mathcal{L}_{grad}$（**grad**），4：$\mathcal{L}_{score}+\mathcal{L}_{attn}+\mathcal{L}_{grad}$（**attn&grad**）の 4 通りの学習方法を比較する．

3.5 評価指標

本実験では，採点精度の評価尺度として quadratic weighted kappa (QWK) [9] を使用する．説明性は，妥当性を根拠箇所の一致度 [2]，忠実性を削除率 [15, 16]，包括性 [17]，十分性 [17] で評価する．採点項目毎にシード値の異なる 10 個のモデルを学習し，その平均を実験結果とする．

根拠箇所の一致度　根拠箇所の一致度は妥当性の評価尺度である．f1 値を用い，モデルが提示した根拠箇所と，真の根拠箇所との一致度合いを測る．図 1 の事例において、真の根拠箇所は「西洋　人」であり、対してモデルが提示した根拠箇所が「人　は　他人」だったとする．このとき，真陽性は 1（人）であり，偽陽性は 2（は　他人），偽陰性は 1（西洋）である．よって，precision は $1/(1+2)=0.33$，recall は $1/(1+1)=0.50$ となり，f1 値は $2\times0.50\times0.33/(0.50+0.33)=0.375$ である．

表 4 十分性（忠実性，↓）の結果

特徴マップ \ モデル	uns	attn	grad	attn&grad
注意重み	0.034	0.037	0.029	0.028
Input X Gradient	0.026	0.031	0.020	0.024
Integrated Gradients	**0.021**	**0.025**	**0.018**	**0.020**
Saliency Map	0.035	0.038	0.024	0.029
Random	0.101	0.116	0.121	0.126

表 5 包括性（忠実性，↑）の結果

特徴マップ \ モデル	uns	attn	grad	attn&grad
注意重み	**0.308**	0.329	0.367	0.381
Input X Gradient	0.267	0.326	0.346	0.363
Integrated Gradients	0.306	**0.367**	**0.378**	**0.418**
Saliency Map	0.239	0.282	0.328	0.328
Random	0.089	0.103	0.106	0.111

削除率　削除率は Serrano ら [15] や Mohankumar ら [16] による忠実性の評価指標である．特徴マップの値が高い単語から降順にマスクしていき，予測ラベルが変化するまでに削除した単語の比率を計算する．特徴マップがモデルの予測過程を反映しているならば，より早い段階で予測が変化すると考えられ，削除率が低いほど忠実度が高いといえる．

先述の通り，自動採点は加点式の情報包含問題である．そのため，得点予測が 0 点の答案は入力単語のマスクによる予測ラベルの変化が生じないと考えられ，そのような答案は評価対象から除外する．

包括性, 十分性　包括性と十分性は Deyoung ら [17] による忠実性の評価指標である．包括性は，根拠箇所が予測にどれほど影響を及ぼすかを表し，大きいほど忠実性が高い．十分性は，根拠箇所が予測を行う上での十分な情報を含むかを表し，小さいほど忠実度が高い．入力 $\boldsymbol{x}$ に対してモデル m が出力するクラス j の予測確率を $m(\boldsymbol{x})_j$，モデルが出力する特徴マップの上位 k% の単語以外をマスクした入力を $\boldsymbol{x}_{\boldsymbol{r}_k}$ とすると，包括性は $m(\boldsymbol{x})_j - m(\boldsymbol{x}\backslash\boldsymbol{x}_{\boldsymbol{r}_k})_j$，十分性は $m(\boldsymbol{x})_j - m(\boldsymbol{x}_{\boldsymbol{r}_k})_j$ で表せる．彼らの実験では，$k=1,5,10,20,50$ の評価値を平均しているが，我々も彼らの実験設定に倣う．また削除率同様に，得点予測が 0 点の答案は評価対象から除外する．

表 6　ある採点項目におけるモデル出力の例

特徴マップ	等案	削除率
Gold	テルは試合に出たのできっちり引退できたが、渡瀬は試合に出ていなく、心のおさまりがつかないため、**やりきれない気持ち**。	–
注意重み	テルは試合に出たのできっちり引退できたが、渡瀬は試合に出ていなく、心のおさまりがつかないため、**やりきれない気持ち**。	0.471
Integrated Gradients	テルは試合に出たのできっちり引退できたが、渡瀬は試合に**出ていなく、心**のおさまりがつか**ないため**、やりきれない**気持ち**。	0.118

4　実験結果

表 1 は損失関数ごとの QWK の結果である．先行研究 [2, 3] と同様に，注意層を教師あり学習することにより，QWK が向上した．また，勾配を教師あり学習することで，同様に QWK が向上した．注意層と勾配の教師あり学習をどちらも行うことによって，どちらか一方を使用する時と比較して更に QWK が向上した．今回の研究は説明性に対する研究であるため，QWK 向上の原因には触れないが，非常に興味深い今後の研究課題である．

妥当性の結果　表 2 は根拠箇所の一致度（f1 値）の結果である．uns ではどの手法も値に優劣がないが，attn や grad，attn&grad では Random を除いたどの特徴マップでも一致度が向上することがわかった．また，attn では注意重みの一致度が，grad では Integrated Gradients の一致度が各々大きく向上していることから，根拠箇所提示手法として活用したい特徴マップと，学習の手法を揃えることで，妥当性をより高めることが可能であることがわかった．

忠実性の結果　表 3，4，5 は削除率，十分性，包括性の結果である．uns の包括性による評価では，注意重みが最も忠実性が高く，それ以外の実験設定，評価指標では，Integrated Gradients が最も忠実性が高かった．また，評価指標により優劣は入れ替わるものの，注意重みは Input X Gradient より忠実性が高かった．ここから，注意重みは Integrated Gradients に劣りつつ，勾配ベースの特徴マップと比較し同等程度に忠実性が高いことがわかった．先行研究では，注意重みは勾配ベースの手法と比較して忠実性に劣ることが報告されている [15, 18, 16, 19] が，今回の実験ではそのような現象は見られなかった．

分析　実験から，注意重みは妥当性が高く，Integrated Gradients は忠実性が高いことがわかった．採点項目ごとに両者の削除率を分析すると，注意重みが Integrated Gradients に大きく負けている採点項目がいくつか見られた．それらの採点項目で何が起きているのか，具体事例を元に分析を行なった．

表 6 に，attn における注意重みと Integrated Gradients によって提示された根拠箇所及び削除率の事例を記した．等案の青色太文字は真の根拠箇所または，モデルが予測した根拠箇所である．注意重みは真の根拠箇所と一致度が高い一方，Integrated Gradients は真の根拠箇所の周辺に着目している事例が散見された．注意重みと比較して Integrated Gradients は削除率が低いことから，モデルは本来着目して欲しい真の根拠箇所ではなく，その周辺単語を元に得点を予測している，つまり，擬似相関を起こしている可能性が考えられる．

5　おわりに

本論文では，自動採点における根拠箇所提示手法の説明性向上を目指し，Mizumoto らによって提唱された根拠箇所提示問題を，Jacovi らによって提唱された説明性を元に再定義した．また，複数の損失関数と複数の根拠箇所提示手法の組み合わせで複数パターンの実験を行い，最も説明性の高い根拠箇所提示手法を検討した．実験の結果，注意重みは最も妥当性が高く，勾配を用いた手法と同程度に忠実性が高いことがわかった．また，Integrated Gradients は最も忠実性が高く，注意重みと同程度に妥当性が高いことがわかった．ここから，自動採点の分野では，説明性の妥当性と忠実性，どちらの観点を重要視するかを元に根拠箇所提示手法を使い分けることが重要であることがわかった．

先行研究では，注意重みを用いた特徴マップは勾配を用いた特徴マップと比較して忠実性に劣ると言われいる [15, 18, 16, 19]．しかし，これらの研究による結果は，「注意重みの忠実性が低下する現象が起こりうる」ことを示すものであり，「注意重みが必ず忠実性が低い」ことを主張するものではないため，本研究の実験結果と矛盾するものではない．

今回，採点項目によっては擬似相関によって注意重みの忠実性が低下することもわかった．今後の方針として，モデル内部の擬似相関を改善する手法を考えていきたい．

謝辞

実際の模試データを提供していただいた学校法人高宮学園代々木ゼミナールに感謝します.

参考文献

[1] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In **Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[2] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic score prediction and justification identification in automated short answer scoring. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 316–325, Florence, Italy, August 2019. Association for Computational Linguistics.

[3] Hiroaki Funayama, Shota Sasaki, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, Masato Mita, and Kentaro Inui. Preventing critical scoring errors in short answer scoring with confidence estimation. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop**, pp. 237–243, Online, July 2020. Association for Computational Linguistics.

[4] 【業界初】代ゼミが「記述式をai採点する現代トレーニング」をリリース. https://www.yozemi.ac.jp/news/press/__icsFiles/afieldfile/2021/05/27/aisaiten_210527.pdf. Accessed: 2022-01-10.

[5] Fei Wang, Rainu Kaushal, and Dhruv Khullar. Should health care demand interpretable artificial intelligence or accept "black box" medicine? **Annals of internal medicine**, Vol. 172, No. 1, January 2020.

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[7] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. **IEEE Transactions on Neural Networks and Learning Systems**, Vol. 32, No. 11, p. 4793–4813, Nov 2021.

[8] 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.

[9] J. Cohen. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. **Psychological bulletin**, Vol. 70 4, pp. 213–20, 1968.

[10] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In **EMNLP**, pp. 230–237, 2004.

[11] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[12] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences, 2017.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In **Proceedings of the 34th International Conference on Machine Learning - Volume 70**, ICML'17, p. 3319–3328. JMLR.org, 2017.

[14] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Learning what makes a difference from counterfactual examples and gradient supervision. **CoRR**, Vol. abs/2004.09034, , 2020.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. Why attentions may not be interpretable? In **Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery amp; Data Mining**, KDD '21, p. 25–34, New York, NY, USA, 2021. Association for Computing Machinery.

A　参考情報

A.1　モデル

注意層　注意機構は Lin ら [**?**] の自己注意を用いる．d_h 次元の隠れベクトルを $\boldsymbol{h_t}$，$d_{out} \times d_h$ 次元の行列を W_{s1}，d_{out} 次元のベクトルを $\boldsymbol{w_{s2}}$ とすると，注意による重み付き平均 $\boldsymbol{h_{attn}}$ は以下のように求められる．

$$a_t = softmax(\boldsymbol{w_{s2}} tanh(W_{s1}\boldsymbol{h_t^T}))$$
$$\boldsymbol{h_{attn}} = \sum_{t=1}^{T} a_t \boldsymbol{h_t}$$

ここで，W_{s1}，$\boldsymbol{w_{s2}}$ は学習パラメータである．

表 7　モデル詳細

単語埋め込み次元数	100
モデルアーキテクチャ	biLSTM
lstm 層数	1
隠れ層次元数	300
FFNN 次元数	300
最適化アルゴリズム	RMSProp
損失関数	交差エントロピー誤差 (得点ラベル) 平均 2 乗誤差 (根拠箇所ラベル)
バッチサイズ	32
エポック数	50
学習率	1.0×10^{-3}

A.2　損失関数

CE を cross entropy loss，N をデータ数，s_p を得点の予測確率，$\hat{s}$ を真の得点，$\hat{r}$ を真の根拠箇所，T を文長，f^{attn} を注意重み，f^{grad} を Integrated Gradients による特徴マップとすると，各損失関数は以下のように表せる．

$$\mathcal{L}_{score} = \frac{1}{N}\sum_{n=1}^{N} CE(s_p^{(n)}, s_g^{(n)})$$
$$\mathcal{L}_{attn} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} (f^{attn}{}_t^{(n)} - \hat{r}_t^{(n)})^2$$
$$\mathcal{L}_{grad} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} (f^{grad}{}_t^{(n)} - \hat{r}_t^{(n)})^2$$