アンサンブル法に基づく深層学習自動採点の不確かさ推定

Ensemble Estimation of Uncertainty in Automated Text Scoring

高橋 祐斗 *1，宇都 雅輝 *1
Yuto Takahashi*1, Masaki Uto*1
*1 電気通信大学
*1The University of Electro-Communications
Email: {takahashi, uto}@ai.lab.uec.ac.jp

あらまし： 近年，人工知能技術を用いた記述式試験の自動採点技術が注目を集めており，特に深層学習を用いた技術の高精度化が目覚ましい．一方で，最先端の深層学習自動採点モデルであっても予測には一定の誤りが含まれている．この課題に対するアプローチの一つとして，自動採点モデルの予測の不確かさを推定することで，予測の誤りを検出する手法が提案されている．本研究では，従来の予測の不確かさ推定手法の性能改善を目指し，深層学習自動採点モデルのアンサンブル学習に基づく新たな予測の不確かさ推定手法を提案する．

キーワード： 記述式試験，自動採点，深層学習，アンサンブル学習，自然言語処理

1　はじめに

近年，記述・論述式試験の採点をコンピュータを用いて自動化する自動採点技術が注目を集めており，特に深層学習を用いた技術の高精度化が進行している．一方で，最先端の深層学習自動採点モデルであっても，予測には一定の誤りが含まれており，このことがハイステークス試験における自動採点実用化の妨げになっている．

この問題を解決する方法の一つとして，自動採点モデルの予測の不確かさを推定することで，予測の誤りを検出する手法が提案されている [1]．予測が不確かである場合，予測が誤っているとみなして人が採点を行うことで，採点のコスト増加を最小化しつつ採点精度を向上できると期待される．一方で，既存手法 [1] の問題点として次の 2 点が挙げられる．1) 一般に自動採点モデルは得点の順序性を考慮するために回帰モデルとして設計されるが，既存手法では分類モデルとして設計したモデルを採用しており，自動採点の精度が低いと予想される．2) 様々な機械学習タスクで有効性が示されている予測の不確かさ推定手法のひとつであるアンサンブル学習に基づく方法 [2] を採用していない．

本研究では，上記の問題を解決するために，新たな予測の不確かさ推定手法を提案する．具体的には，回帰モデルとして設計した深層学習自動採点モデルを構築し，そのモデルのアンサンブル学習により予測の不確かさを推定する手法を提案する．なお，提案自動採点モデルのベースには，従来手法と同じく BERT（Bidirectional Encoder Representations from Transform）を利用する．本研究では，実データ実験を通して，既存手法 [1] と提案手法の性能を評価する．

2　先行研究

先行研究 [1] で提案された予測の不確かさ推定手法では，分類モデルとして設計した BERT を自動採点モデルとして採用している．このモデルでは，答案 $\boldsymbol{x}$ を BERT に入力して得られる文章の分散表現ベクトル $\boldsymbol{h}$ を線形変換することで得点段階数と同長のベクトルを $\boldsymbol{u} = \boldsymbol{W}_c\boldsymbol{h} + \boldsymbol{b}_c$（ここで $\boldsymbol{W}_c, \boldsymbol{b}_c$ はパラメータ）で求め，それを Softmax 関数に与えることで，各得点 k に対する分類確率を $P_k = \exp(u_k)/\sum_{i=1}^{K}\exp(u_i)$（ここで K は得点段階数，u_k はベクトル $\boldsymbol{u}$ の k 番目の要素）と求める．このモデルでは，答案 $\boldsymbol{x}$ に対する予測得点を $\hat{y}_c = \underset{k}{\mathrm{argmax}} P_k$，その予測の不確かさを $U_{prob} = -P_{\hat{y}_c}$ で定義する．

3　提案手法

3.1　深層学習自動採点モデル

提案手法では，分類モデルよりも一般に高精度な予測が可能な回帰モデルとして BERT を設計する．モデルの概念図を図 1 に示す．このモデルは，得点とその得点に対する分散を出力する．具体的には，答案 $\boldsymbol{x}$ を BERT に入力して得られる文章の分散表現ベクトル $\boldsymbol{h}$ を次式で変換することで，$\boldsymbol{x}$ の得点 $\hat{y}_{reg}$ と $\hat{y}_{reg}$ に対する分散の対数値 $\log\hat{\sigma}^2_{reg}$ を計算する．

$$\hat{y}_{reg} = \mathrm{sigmoid}(\boldsymbol{W}_s\boldsymbol{h} + \boldsymbol{b}_s) \tag{1}$$

$$\log\hat{\sigma}^2_{reg} = \boldsymbol{W}_v\boldsymbol{h} + \boldsymbol{b}_v \tag{2}$$

ここで，$\boldsymbol{W}_s, \boldsymbol{W}_v, \boldsymbol{b}_s, \boldsymbol{b}_v$ は重みとバイアスを表すパラメータ，sigmoid はシグモイド関数を表す．このモデルでは，予測の不確かさを $U_{var} = \hat{\sigma}^2_{reg}$ で評価できる．

モデルの学習は次式の損失関数 L_r を最小化することで行われる．

$$L_r = \sum_{j=1}^{J}\{\frac{\|y_j - \hat{y}_j\|^2}{2\hat{\sigma}^2_j} + \frac{1}{2}\log\hat{\sigma}^2_j\} \tag{3}$$

ここで，$\hat{y}_j$ と $\hat{\sigma}^2_j$ は答案 $\boldsymbol{x}_j$ に対する得点と分散の予測値，y_j は答案 $\boldsymbol{x}_j$ の真の得点である．

3.2　予測の不確かさのアンサンブル推定

本研究では，上記のモデルのアンサンブル学習に基づく新たな予測の不確かさ推定手法を提案する．提案手法

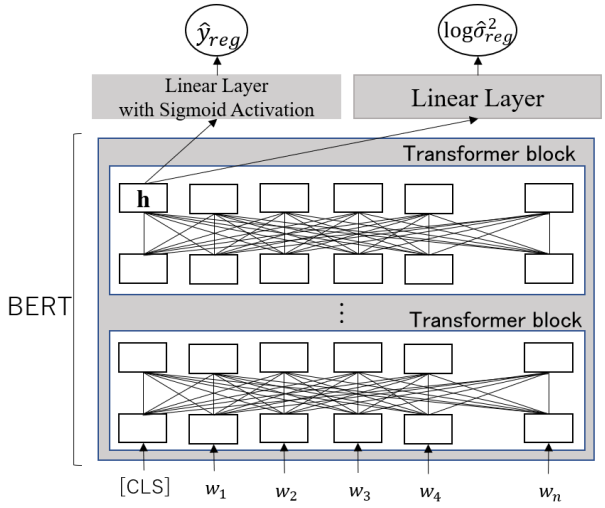

図 1　本研究で開発した自動採点モデルの概念図

では，訓練データからランダムに抽出した部分データを用いて自動採点モデルを訓練する操作を繰り返すことで，異なるパラメータを持つモデルを複数用意し，それらの予測結果を統合することでアンサンブルを行う．具体的には，Lakshminarayanan ら [2] のアンサンブル手法に基づき，答案 $\boldsymbol{x}$ を複数のモデルに入力して得られる出力を用いて，$\boldsymbol{x}$ の得点 $\hat{y}_{mul}$ とそれに対する分散 $\hat{\sigma}^2_{mul}$ を次式で求める．

$$\hat{y}_{mul} = \frac{1}{R}\sum_{r=1}^{R}\hat{y}_r \tag{4}$$

$$\hat{\sigma}^2_{mul} = -\hat{y}^2_{mul} + \frac{1}{R}\sum_{r=1}^{R}(\hat{\sigma}^2_r + \hat{y}^2_r) \tag{5}$$

ここで，R はアンサンブルするモデルの数であり，$\hat{y}_r$ と $\hat{\sigma}^2_r$ は r 番目のモデルに答案 $\boldsymbol{x}$ を入力して得られる得点と分散の予測値である．提案手法では，$\hat{y}_{mul}$ に対する不確かさは $U_{mul} = \hat{\sigma}^2_{mul}$ で定義される．

4　評価実験

ここでは，実データ実験により，提案手法と既存手法 [1] における自動採点の精度と予測の不確かさ推定の性能を評価する．本実験では，短答記述式自動採点に利用されるデータセットを使用した [3]．このデータセットは 6 つの短答記述問題で構成され，各問題は複数の採点項目で得点付けされている．採点項目数は合計で 21 であり，課題あたりの受験者数の平均は 2100 人である．

本実験では，このデータセットを用いて，採点項目ごとに交差検証法でモデルの性能評価を行った．採点精度の評価には，Root Mean Squad Error（RMSE）と相関係数を利用し，予測の不確かさ推定の性能評価には Receiver Operating Characteristic（ROC）曲線の曲線下面積（Area Under the Curve; AUC）を用いた．ROC-AUC は，値が 1 に近くなるほど，予測の不確かさが適切に推定できていることを意味し，0.5 に近いほど不確かさ推定が不適切であることを意味する．

表 1　実験結果

	従来手法	提案手法	提案手法 +
RMSE	0.56*	0.53*	**0.50**
相関係数	0.89*	0.90*	**0.91**
ROC-AUC	0.88*	0.88*	**0.91**

本実験では，次の 3 つの手法について性能を比較した．

- **従来手法**：分類 BERT に基づく方法
- **提案手法**：回帰 BERT に基づく方法
- **提案手法＋**：回帰 BERT によるアンサンブル法

なお，アンサンブル法で利用するモデル数は 5 つとした．

また，既存手法と提案手法の有意差を確認するために，提案したアンサンブル手法を基準として他の手法との対応のある t 検定を行った．

実験結果を表 1 に示す．表には，データセット中の 21 項目について得られた結果の平均のみを記している．また，表中の「*」は提案モデル + の平均性能と有意水準 5% で有意差が認められたことを意味する．

表より，予測精度の観点では，本研究で開発した回帰モデルとして設計した BERT が，従来の分類モデルとして設計した BERT よりも高精度であり，提案アンサンブル法が全手法の中で最も高精度であったことがわかる．さらに，ROC-AUC から，予測の不確かさ推定の観点でも，提案アンサンブル手法が他の手法よりも有意に性能が高いことが確認できる．

5　まとめと今後の課題

本研究では，アンサンブル学習に基づく新たな予測の不確かさ推定手法を提案し，比較実験を行った．紙面の都合上割愛したが，本研究では上記の手法に加え「分類モデルとして設計した BERT のアンサンブル」，「ガウス過程回帰」，「trust score」，「MC Dropout に基づくアンサンブル」などの手法も実装し，短答記述式だけでなく小論文評価のデータセットでも評価実験を行い，提案手法の有効性を確認した．

謝辞: 本研究では、国立情報学研究所の IDR データセット提供サービスにより国立研究開発法人理化学研究所から提供を受けた「理研記述問題データセット」を利用した．

参考文献

[1] Hiroaki Funayama, Tasuku Sato, Yuichiroh Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. Balancing cost and quality: An exploration of human-in-the-loop frameworks for automated short answer scoring. In *International Conference on Artificial Intelligence in Education*, pp. 465–476. Springer, 2022.

[2] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, Vol. 30, , 2017.

[3] 理化学研究所（2020）：理研記述問題採点データセット．国立情報学研究所情報学研究データレポジトリ． データセット：https://doi.org/10.32130/rdata.3.1.