ベイジアンニューラルネットワークを用いた記述式答案の自動採点

加藤　嘉浩
ベネッセ教育総合研究所
y-kato@mail.benesse.co.jp

1　はじめに

2022 年度より施行される高等学校の新学習指導要領における重点項目は，思考・判断・表現の育成である．これらを育成するためのアセスメントのひとつに記述式テストが挙げられる．具体的には，大学入学共通テストの試行調査に見られたような複数の資料を整理・読解し，題意に即して記述解答するテストである．本研究で扱う記述式テストは，明確な採点基準があり，エッセイや正解が無いオープンエンドの記述式テストとは異なる．大学入学共通テストにおける記述式問題の導入は見送られたが，新学習指導要領の重点項目に変更はなく，大学の個別選抜試験や AO 入試，学校現場において（採点基準が明確な）記述式テストの需要が高まることが予測される．しかし，記述式テストは，採点および指導に長い時間を要し，人材の確保および育成が困難なため効率化が望まれる．

そこで，機械学習手法を用いて自動採点を行うことが考えられ，これまでに多くの研究がされている．例えば，エッセイの自動採点は，英語では e-rater[1]，日本語では Jess[2] が挙げられる．短文では，中島 [3] が SVM を用いて実運用上の問題を検討している．近年，深層学習を用いた短文の自動採点モデルが数多く発表されている．寺田ら [4] は単語の依存関係を用いた畳み込みニューラルネットワークを用い高精度な結果を示している．水本ら [5] は，解答データの採点基準に該当する箇所にアノテーションを施し，注意機構（Attention）と LSTM を用いて各採点基準の得点と全体得点を予測するモデルを提案している．また，得点予測だけでなく，Attention の可視化を通して学習者へのフィードバックを検討している．王ら [6] は，Attention に採点基準の文章を用い得点予測，学習者へのフィードバックを検討している．竹谷ら [7] は，実際のシステム運用を念頭に置き，多くの前処理とキーフレーズ比較により非常に高精度な結果を示している．以上のように，日本国内においても多くの自動採点研究が行われており，高精度かつ有用性の高い知見が示されている．特に，深層学習を用いることで高精度に採点予測できることが報告されている．したがって，本研究でも深層学習を用いることとする．

深層学習特有の問題点として，高い確信度で誤った予測をしてしまう点が挙げられる．教育現場において，誤った採点結果を学習者に返却することは避けねばならない．特に，正答であるはずの回答を誤答と判断した場合，学習者の混乱を招き，意欲低下の原因になりかねない．これは深層学習がデータに対し過剰適合していることや確率モデルではないためデータの不確実性を考慮できないことが原因である．これに対し，深層学習をベイズ化したベイジアンニューラルネットワーク（Bayesian Neaural Network，BayesianNN）が有用であると考えられる．BayesianNN を用いることにより，予測結果を確率分布として得られ，予測の不確実性を定量的に評価できる．本研究では，自動採点に BayesianNN を適用し，その有効性を検証する．

BayesianNN の有効性を検証するため実データを用いて実験を行った．実験では，研究開発中の記述式テストから回答制限字数が異なる 3 つの問題を使用し，BayesianNN および自動採点で用いられる Attention モデルを用いて予測精度を比較した．また，BayesianNN の予測分布の分散と予測の平均二乗誤差（MSE）を比較した．

2　関連研究

深層学習を用いた自動採点は，人手の採点結果を教師データとした文書分類問題として解ける．代表的なモデルとして，畳み込みニューラルネットワーク（Convolutional Neaural Network，CNN）と長・短期記憶（Long Short-Term Memory，LSTM）が挙げら

れる．また，自己注意機構（Self-Attention）を追加することで，予測性能が向上することが知られている [8, 10]．

2.1 ベイジアンニューラルネットワーク

BayesianNN は，学習データ D が与えられた時のネットワークの重み w を誤差逆伝搬時にサンプリングして求める．このとき，真の事後分布 $p(w|D)$ が解析的に求められないため，計算可能な変分確率分布 $q_\theta(w|D)$ を導入し，真の事後分布との KL ダイバージェンス（KLD）を最小にする最適化問題に帰着させる．

Shridhar ら [9] は，確率的変分法によるベイジアン畳み込みニューラルネットワーク（BayesianCNN）を提案しており，CNN を基にした代表的な画像分類モデルをベイズ化し分類精度を向上させている．Shridhar らは，変分事後確率分布を $q_\theta(q_{ijhw} \mid D) = \mathcal{N}(\mu_{ijhw}, \alpha_{ijhw}\mu_{ijhw}^2)$ と定義し，Local Reparameterization Trick を CNN に適用し，w をサンプリングする代わりに式 (1) に示す活性値 b をサンプリングをする．b_j は，平均 μ_{ijhw}，分散 $\alpha_{ijhw}\mu_{ijhw}^2$ の関数であり，畳み込み演算の際に平均と分散を個々に算出している．

$$b_j = A_i * \mu_i + \epsilon_j \odot \sqrt{A_i^2 * (\alpha_i \odot \mu_i^2)} \qquad (1)$$

ここで，i, j は入力と出力層，h, w はフィルターの高さと幅を示す．$\epsilon_j \sim \mathcal{N}(0, 1)$，$A_i$ は受容野，$*$ は畳み込み，$\odot$ は各コンポーネントの積を示す．

3 実験

3.1 データ

実験に使用した記述式テストは，明確な採点基準があり，今回はスペースの都合により観点 4 つを抜粋した．採点者 2 人の合議によりすべての回答を採点した．表 1 に，すべての解答データを形態素解析器 JUMAN[1] により分かち書きした際の単語数の平均（標準偏差），語彙数を示す．また，解答に含まれるすべての文字列を評価対象とするため，ストップワードの除去や表記揺れの修正は行わなかった．図 1, 2, 3 に各問題の得点分布を示す．

[1] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN

表 1: データの概要

問題	単語数	語彙数	データ数
問題 1	74.82(19.71)	1254	421
問題 2	133.87(21.85)	908	498
問題 3	271.21(92.74)	2931	781

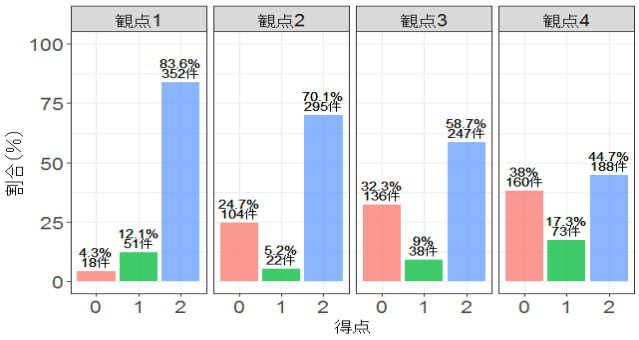

図 1: 問題 1 の得点分布

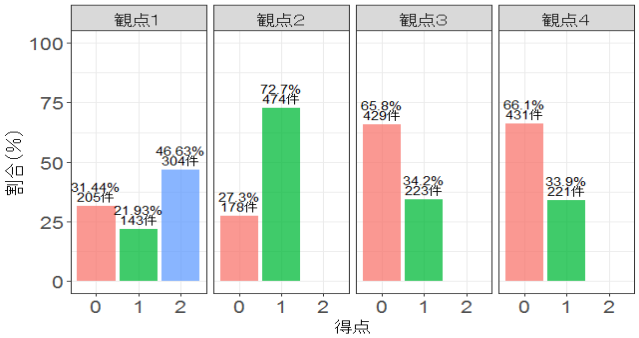

図 2: 問題 2 の得点分布

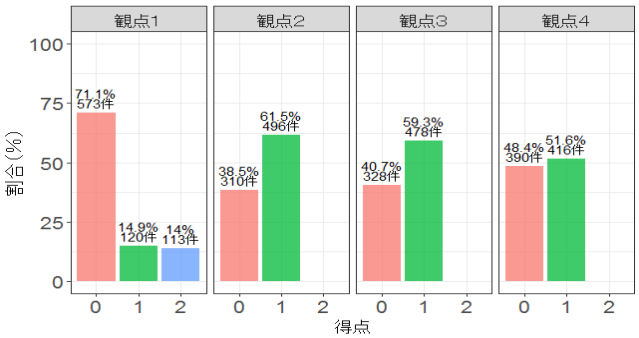

図 3: 問題 3 の得点分布

3.2 方法

実験には，CNN をベイズ化した BayesianCNN，CNN に self-Attention を追加した CNN_Attention，LSTM に self-Attention を追加した LSTM_Attention を使用した．すべてのモデルの入力は，解答データを JUMAN により分かち書きしたものを使用した．単語埋め込みの次元数は 200 とした．BayesianCNN および

CNN_Attention のフィルタの高さ h を 3，4，5 とし，フィルタの数は 200 とした．BayesianCNN は，畳み込み層の出力に対し活性化関数 softplus（$\beta = 1$），max-pooling を適用し，全結合層，Bayesian Linear 層を経て softmax により予測結果を得る．CNN_Attention は，畳み込み層の出力に対し活性化関数 ReLU，Attention による pooling，全結合層で Dropout（$p = 0.5$）を適用し，Linear 層を経て softmax により予測結果を得る．LSTM_Attention は隠れ層の次元数を 200 とし，Atention を適用し，Linear 層を経て softmax により予測結果を得る．

すべての実験において，各問題の観点別に 5-fold 交差検証を行い，エポック数を 100，バッチサイズを 16，学習率を 0.001，最適化手法は Adam を用いた．評価指標は，Quadratic Weighted Kappa（qwk）を用いた．

4 結果

表 2 に，モデルと各観点に対する qwk の交差検証の平均（分散）を示す． 表 2 から，BayesianCNN は問題 1 の観点 1，2，4，問題 2 の観点 2，3 において最も精度が高い結果となった．CNN_Attention は，問題 1 の観点 3，問題 2 の観点 1，4，問題 3 の観点 1，3，4 で最も高い精度だった．LSTM_Attention は，問題 2 の観点 3，問題 3 の観点 2 で最も高い精度となった．各問題における各モデルの qwk の平均から，問題 1 は BayesianCNN，問題 2，3 は CNN_Attention が精度が最も高い精度だった．また，BayesianCNN よりも CNN_Attention の方が高精度な結果となった観点が多い結果となった．特に，回答字数が最も長い問題 3 において CNN_Attention が高精度な結果を示している．

BayesianCNN はデータの不確実性を考慮したモデルのため，予測分布の分散が計算できる．図 4 に，BayesianCNN のテストデータに対する予測分布の分散と MSE を問題別に示す．図中の点は，交差検証のテストデータに対する予測分布の分散と MSE を示す．図 4 から，予測分布の分散が高くなるにつれ，MSE も高くなっている．問題 1 の観点 3，4，問題 2 の観点 1，問題 3 の観点 1，2 など qwk の値が低いデータは，予測分布の分散と MSE が高い結果となった．

5 おわりに

本研究では，データの不確実性を考慮した BayesianCNN を記述式テストの自動採点に適用した．実験では，CNN および LSTM に Self-Attention を追加したモデルを比較手法に用い，回答字数が異なる 3 つの記述式テストの回答データに対する予測精度を交差検証で評価した．回答字数が最も短い問題 1 では，BayesianCNN が最も精度が高い結果となったが，その他の問題では CNN_Attention が高精度な結果となった．言語処理では，CNN よりも文脈を考慮できる LSTM を基にしたモデルの方が適していると考えられるが，今回の実験では CNN を基にしたモデルの方が高精度な結果になった．

BayesianCNN の予測分布の分散を算出し，予測の確信度と予測誤差の関係性を定量的に評価した．分散が大きくなるにつれ MSE が大きくなる結果となったが，誤った予測の原因究明やモデルへの反映には至っていない．

実験で使用した記述式テストの作問担当者らは，記述式テストで思考・判断・表現を測定するためには，問題 3 程度の字数（もしくはより長く）が望ましいと考えている．しかし，本研究で示した予測精度では，問題 3 のような長文の自動採点の実用化は無理があり，既存手法においても実用に耐えるモデルは見当たらない．今後は，誤った予測の原因の定量的評価，誤った予測を未然に防止可能なモデルの構築，そして長文データの得点を高精度に予測するモデルの構築を目指す．

参考文献

[1] Yigal Attali and Jill Burstein. *Automated essay scoring with e-rater v.2.* Journal of Technology, Learning, and Assessment, 2006.

[2] 石岡恒憲, 亀田雅之. コンピュータによる小論文の自動採点システム *Jess* の試作. 計算機統計学, Vol.16, No.1, pp.3-19, 2003.

[3] 中島功滋. 短答式記述答案の採点支援ツールの開発と評価. 言語処理学会, 第 17 回年次大会 発表論文集, pp.611-614, 2011.

[4] 寺田凛太郎, 久保顕大, 柴田知秀, 黒橋禎夫, 大久保智哉. ニューラルネットワークを用いた記述式問題の自動採点. 言語処理学会 第 22 回年次大会 発表論文集, pp.370-373, 2016.

表 2: 実験結果

問題	モデル	観点 1	観点 2	観点 3	観点 4	平均
問題 1	Bayesian_CNN	**0.962(0.000)**	**0.932(0.002)**	0.768(0.009)	**0.414(0.005)**	**0.769**
	CNN_Attention	0.937(0.002)	**0.932(0.002)**	**0.841(0.003)**	0.348(0.018)	0.765
	LSTM_Attention	0.923(0.004)	0.433(0.056)	0.495(0.025)	0.212(0.004)	0.516
問題 2	Bayesian_CNN	0.735(0.002)	**0.936(0.001)**	**0.911(0.000)**	0.832(0.001)	0.854
	CNN_Attention	**0.866(0.002)**	0.925(0.000)	0.905(0.002)	**0.850(0.002)**	**0.887**
	LSTM_Attention	0.741(0.005)	0.935(0.000)	**0.911(0.000)**	0.839(0.003)	0.857
問題 3	Bayesian_CNN	0.698(0.002)	0.498(0.003)	0.716(0.000)	0.869(0.001)	0.695
	CNN_Attention	**0.760(0.002)**	0.657(0.004)	**0.758(0.001)**	**0.906(0.002)**	**0.770**
	LSTM_Attention	0.248(0.049)	**0.820(0.001)**	0.703(0.001)	0.826(0.002)	0.649

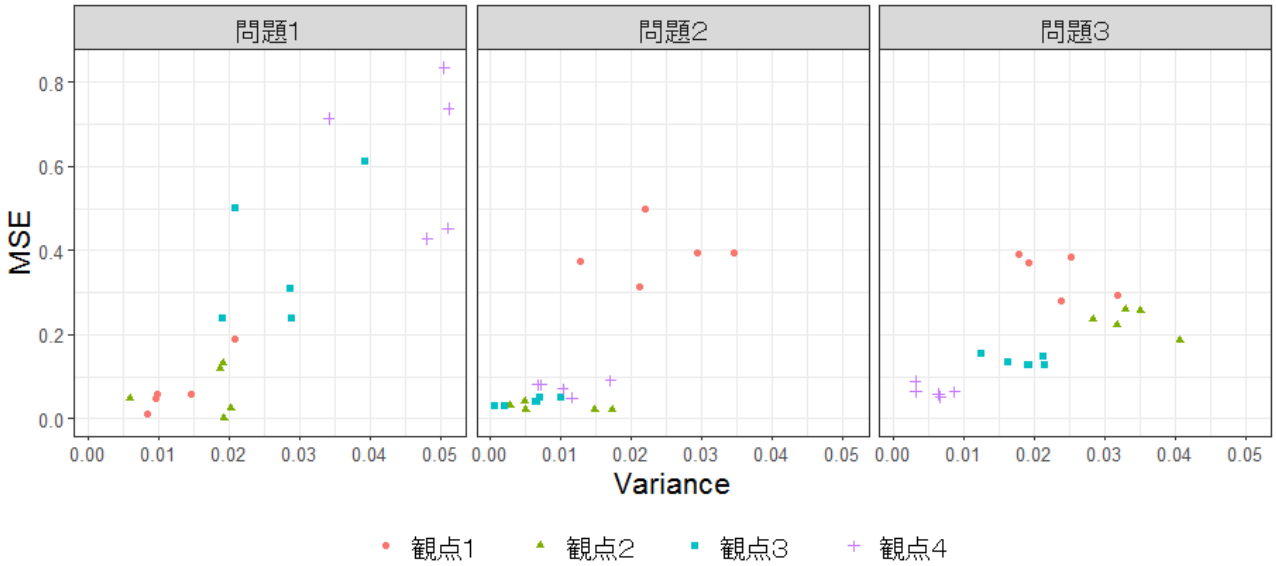

図 4: 予測分布の分散と MSE

[5] 水本智也, 磯部順子, 関根聡, 乾健太郎. 採点項目に基づく国語記述式答案の自動採点. 言語処理学会第 24 回年次大会 発表論文集, pp.552-555, 2018.

[6] 王天奇, 井之上直也, 水本智也, 大内啓樹, 乾健太郎. 採点基準を利した記述式答案の動採点. 言語処理学会 第 25 回年次大会 発表論文集, pp.450-453, 2019.

[7] 竹谷謙吾, 高井浩平, 清水杏奈, 早川純平, 森康久仁, 須鎗弘樹. 大規模実データにおける記述式問題自動採点システムの検証. 言語処理学会 第 25 回年次大会 発表論文集, pp.880-881, 2019.

[8] Tao Shen and Tianyi Zhou and Guodong Long and Jing Jiang and Shirui Pan and Chengqi Zhang. *DiSAN: Directional Self-Attention Network for RNN/CNN-free Language Understanding.* The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), pp.5446–5455, 2018

[9] Shridhar Kumar,Laumann Felix,Liwicki Marcus. *A Comprehensive guide to Bayesian Convolutional Neural Network with Variational Inference.* arXiv preprint arXiv:1901.02731, 2019.

[10] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, Daniel Rueckert. *Attention gated networks: Learning to leverage salient regions in medical images.* Medical Image Analysis, Vol.53, pp.197–207, 2019.