7C-03

軽量なOCRモデルの予測確信度と言語モデルを利用した日本語手書き文字認識結果の事後訂正

大城 海斗[†1]　仲宗根 慎太[†2]　三上 威[†3]　松本 涼[†2]　真嘉比 愛[†1]
ちゅらデータ株式会社[†1]　琉球大学大学院理工学研究科[†2]　アーリース情報技術株式会社[†3]

1. はじめに

光学文字認識(Optical Character Recognition; OCR)は、画像として取り込まれた書類を文字コードへ変換する技術であり、データ入力作業の効率化や書類の検索性向上などの重要な役割を担っている。認識対象の書類には印字だけでなく手書き文字を含む場合がある。手書き文字は書き手によって癖が異なるため、認識の難易度が印字に比べて高く、誤認識しやすい傾向にある。

深層学習は多くの画像認識タスクにおいて大きな成功を収めており、OCR を実現するアプローチとしても注目されている。中でもリアルタイム推論を目的とした軽量なモデルの開発が盛んだが、軽量化する代償として認識性能が低下することが報告されている[1]。本稿では、軽量なOCR モデルの認識性能が低下するという課題に対し、OCR モデルの予測確信度と言語モデルを利用した認識結果の事後訂正手法を検討した。

2. 文字認識の関連技術

近年の深層学習を用いたOCRモデルに、文字毎へ分割せずに文字列単位で認識を試みる Seq2Seq ベースモデルがある。Seq2Seq ベースモデルにおいて、エンコーダに ResNet を、デコーダに Transformer を用いる取り組みが行われている[2]。このような構造が大きいモデルは、認識性能が高いというメリットがある一方で、モデルサイズの肥大化に伴う学習にかかるコストの増加や推論速度の低下といったデメリットが生じる。PP-OCR[1]では、エンコーダに MobileNetV3 や中間層のユニット数を少なく設定したRNNといった軽量なネットワークを採用することで、推論時のみならず学習時のコストを低減した。一方で、ResNet34 を用いた大規模モデルとの精度比較では、軽量化したモデルのF-scoreは大規模モデルよりもやや低かったと報告されている。

3. 文字認識結果の誤りパターン

日本語手書き文字を PP-OCR で認識させた結果、字形が類似した文字への誤りが誤認識の 67%を占めることがわかった。類似した文字に誤認識した例として「士」を「土」と誤認識した図1の「佐久市長士呂」を示す。「士」に対する予測確信度の分布（図 2）を可視化すると、正解文字である「士」が2番目に高い予測確信度であることが確認できた。類似した文字へ誤認識した他の例も同様に、誤った文字が1番目に確信度が高く、2番目から5番目に高い予測確信度で正しい文字が現れる傾向にあることがわかった。

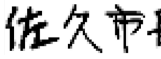

図 1 「士」を「土」と類似した文字へ誤った例

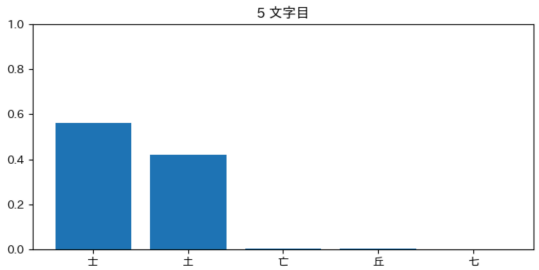

図 2 図1の5文字目の予測確信度の分布

4. 文字認識結果の事後訂正手法

誤認識の半数以上を占めている類似した文字への誤りを改善できればOCRモデルの認識性能を向上させることが可能だと考えられる。本稿では、類似した文字への誤りを訂正する手法を検討する。

OCR の誤り訂正手法として、竹内ら[3]は統計的言語モデルを用いて、訂正箇所検出、訂正候補生成、訂正候補選択の手順で誤りを訂正するシステムを提案した。この訂正手法では各処理を行うために、文字あるいは単語の出現頻度や類似度、混同確率といった情報を保持する必要があり、訂正手法の軽量化の妨げとなることが懸念される。そこで、本稿ではOCRモデルで得られる予測確信度と言語モデルを用いて訂正候補を予測確信度が高い文字から選択することで、文字の出現頻度などの情報を保持せずに訂正できる手法を検討した。長さ n の訂正候補文字列

s の i 番目の文字の予測確信度を $conf_{s,i}$ 、訂正候補文字列に対する言語モデルのスコアを lm_score_s 、言語モデルのスコアの重みを α としたときに式(2)を満たす訂正候補文字列 s^* を訂正後の文字列とする。ただし、 S_τ を$conf_{s,i}$が「確信度の最大値 $\times\tau$ 以上」となる訂正候補文字列の集合とし、予測確信度が著しく低い文字を訂正候補文字列の集合から除外する。

$$fix\_score_s = \frac{1}{n}\sum_{i=1}^{n} conf_{s,i} + \alpha \cdot lm\_score_s \quad (1)$$

$$s^* = \underset{s \in S_\tau}{\operatorname{argmax}}\ fix\_score_s \quad (2)$$

5. 認識精度および軽量さの評価

文字認識にバックボーンを MobileNetV3、バッチサイズを 64、学習エポック数を 20、初期の学習率を 0.0001 に設定した PP-OCR を用いた。PP-OCR の学習及び検証用のデータセットは MeCab 辞書[7][8]から取得した単語や句を元に、手書き文字データベースである ETL データベース[4]と手書き文字風フォントを用いて人工的に生成した（表 1）。言語モデルには日本語 Wikipedia データを利用して学習させた KenLM[5]と Faster RNNLM[6]を用い、それぞれが訂正にどの程度貢献したかについて、訂正できたデータ数と文字誤り率（CER）で評価した。CER は正解文字列長を n 、レーベンシュタイン距離における文字の挿入数を i 、置換数を r 、削除数を d としたとき$(i+r+d)/n$で算出される。また、軽量さの評価として、言語モデルのサイズおよび訂正に要する時間を計測した。

無作為に検証データセットを 2,000 件選び、式(1)(2)中のパラメータ α 、τ を変化させて KenLM と Faster RNNLM のそれぞれで事後訂正を行った結果、表 2 に示すパラメータで最も誤りを訂正することができた。ただし、Faster RNNLM の class 数は 15,456、隠れ層のサイズは 320 とした。KenLM は Faster RNNLM よりも誤りを多く訂正できた一方、3.08GiB というモデルサイズは PP-OCR の 3.0MB と比較すると非常に大きく、軽量化への課題が残る。Faster RNNLM は、学習コーパスに依存して class 数が多くなっているため、日本語の文字種に限定したモデルを設計することで、さらなる軽量化と高速化が図れると考えられる。

	学習データセット（文）	検証データセット（文）
ETL データベース	1,930,755	482,688
DF てがき速 Std W3	2,658,468	664,616
全児童フォント	2,658,468	664,616
手書き屋本舗 TA へたれ R	2,658,468	664,616
手書き屋本舗 TA 礼筆	2,658,468	664,616

表 1 データセットの内訳

言語モデル	α	τ	正解数	誤り数	平均 CER	モデルサイズ (GiB)	平均所要時間(ms)
訂正前	−	−	1,879	121	0.0158	−	−
KenLM	0.05	0.05	1,916	84	0.0123	3.08	0.179
Faster RNNLM	0.02	0.05	1,887	113	0.0154	0.04	1112.185

表 2 事後訂正手法の評価

6. おわりに

本稿では、文字認識結果の事後訂正手法について検討し、OCR モデルの予測確信度と言語モデルを用いて訂正候補を選択する手法の有効性を示した。今後の課題として、言語モデルの軽量化と精度の改善に取り組む。

引用文献

[1] Y. Du, C. Li, R. Guo et al., “PP-OCR: A Practical Ultra Lightweight OCR System,” arXiv:2009.09941 [cs], 2020.

[2] W. Hu, X. Cai, J. Hou, S. Yi, and Z. Lin, “GTC: Guided Training of CTC towards Efficient and Accurate Scene Text Recognition,” AAAI, vol. 34, no. 07, pp. 11005-11012, 2020.

[3] 竹内孔一, 松本裕治, “統計的言語モデルを用いた OCR 誤り訂正システムの構築,” 情報処理学会論文誌, vol. 40, no. 6, pp.2679-2689, 1999.

[4] 電子技術総合研究所, Japanese Technical Committee for Optical Character Recognition, “ETL 文字データベース,” 1973-1984.

[5] K. Heafield, “KenLM: Faster and Smaller Language Model Queries,” Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 187-197, 2011.

[6] T. Mikolov, “Statistical Language Models Based on Neural Networks,” Presentation at Google, Mountain View, 2nd, April 2012.

[7] 佐藤敏紀, 橋本泰一, and 奥村学. “単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討,” 言語処理学会 第 23 回年次大会 発表論文集, pp.875-878, 2017.

[8] 工藤拓, “MeCab: Yet Another Part-of-Speech and Morphological Analyzer,” https://taku910.github.io/mecab/, 2006.

Post-correction of Japanese handwritten character recognition results using prediction confidence of lightweight OCR model and language model

†1 Chura DATA Inc.
†2 University of the Ryukyus
†3 ALIS INFORMATION TECHNOLOGIES, LTD.