ISS-SP-041

部首に注目したDeep Leaningによる くずし字の認識を用いた日本古典籍の解析

中屋 悠資† 鈴木 海渡† 宮崎 智†† 大町 真一郎††

† 東北大学 工学部 電気情報物理工学科　　†† 東北大学 工学研究科

1. はじめに

1000 年以上前の日本では現代の仮名とは異なり，「くずし字」という日本語の筆記体で読み書きが行われていた．しかしながら，1900 年の以降，通常の学校教育でそれらを学ぶことはできず，くずし字を読み書きしうる人口も非常に減少している．

過去の日本の文化，風習などを理解するには古代に記された文章や本を参照することが有効であるが，国書総目録[1]に記されるいわゆる古文書は，1867 年以前のものでも 170 万冊以上あり，未登録のものを合わせるとその数は推計 300 万冊以上にものぼる．そこには現代社会にも有用な情報などがあるためにデジタル化が進められているが，たとえそれらが一般に広く閲覧可能となった場合においても，くずし字で書かれた文章を理解し，情報を得ることは専門家を除くと容易ではない．

そこで本研究では，日本の古文書を専門知識の有無にかかわらずそれらを容易に理解するべく，深層学習（Deep Leaning）によってくずし字によって発生する問題の解決を試みた．

2. くずし字

くずし字を扱う上での問題点として，まず文字が重なっていることが挙げられる．このことから，一文字一文字の境界を取ることが非常に困難になっている．次に，古文書ごとに行間などのレイアウトが異なり，同じ文書内においてもそれらは不規則に変化しているという問題がある．さらには，異なる文字の場合でもくずし字の場合見た目のみで判断することは困難であり，その文脈などから判断する必要がある．これらの問題に加え，古典籍におけるくずし字のデータセットの数は限られており，中にはデータが 1 つのみという漢字なども存在する．そこで本論文では，次節にて提案する「部首」に着目した手法でそれらの問題の解決し，元の漢字を推察しようと試みた．

3. 研究手法

本研究では，深層学習を用いてくずし字画像に含まれる部首 ― 偏(へん）と旁(つくり）を推定する．くずし字画像は，Kuzushiji_Kanji(KMNIST と同時公開のデータセット, (https://github.com/rois-codh/kmnist) [2]から，データ数に偏りがあり，かつ共通の部首を含む 15 個のくずし字を選択した．具体的には，「何河阿拾結紅江波彼抄砂沙紗位泣」で，13 種類の部首を含む．

深層学習には畳み込みニューラルネットワーク（CNN）を用いた．図 1 に各レイヤーの出力を示す．

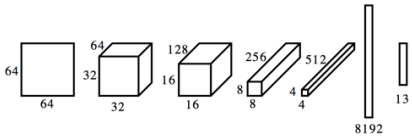
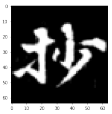


図 1．CNN の各レイヤーの出力．入力は 64 画素四方,各レイヤーは畳み込み, バッチ正規化, Max Pooing で構成した．

実験では，「抄紗」以外の 13 文字を学習データ(画像数 1822 枚)とし，「抄紗」の 2 文字をテストデータ(画像数 23 枚)として用いた．実験の結果，その漢字が含む部首をすべて当てられた場合を正解としたときの認識率は 8.7%となり，テストデータの部首をほとんど認識することができなかった．以下図 2 に，「抄」という文字で認識が成功した例，図 3 に認識に失敗した例を示した．



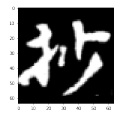


図 2．認識が成功した例（左）

図 3 扌(手偏)の認識ができなかった例(右)

今回の実験結果では，偏の認識率が非常に悪かったが，これは右側の旁の部分が結果に影響を与えている可能性があり，本研究で用いたような単純な CNN では，認識が難しいと考えられる．

4. 結論・今後の展望

本研究では，部首に注目した深層学習によってデータセットに含まれる漢字のインバランスさを解決し，くずし字を解読することを目標とした．しかし，今回用いたような単純な CNN では，十分な精度で認識することができなかったため，今後は部首の位置に関する情報などを事前に与えて，物体検出の手法の利用などを検討したい．

参考文献

[1] 岩波書店 『国書総目録』. 岩波書店, 2002

[2] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep Learning for Classical Japanese Literature arXiv: 1812. 01718, 2018