G-003

深層学習を用いた日本古典文学くずし文字識別
―現代人の模写によるくずし文字データセット拡張の試み―
Deep Neural Network for the Recognition of Classical Japanese Characters

我妻 伸彦† 工藤 雅‡ 駒井 丈瑠‡
Nobuhiko Wagatsuma Miyabi Kudo Takeru Komai

1. はじめに

深層畳み込みニューラルネットワーク(Deep Convolutional Neural Network, DCNN)の発展により、計算機の画像や文字に対する識別能力が極めて向上した[1][2]。しかし、DCNN がヒトに匹敵する高い識別能力を獲得するためには、大規模な学習データが必須となる。本研究では、限られたデータ量である古典文学くずし文字データセット[3]を拡張し、DCNN へと適用した。具体的には、現代人が模写した古典くずし文字を用いて、AlexNet[1]に基づく DCNN を学習させた。模写により拡張された学習データセットは、DCNN のくずし文字識別能を向上させた。

2. 提案手法

2.1 深層畳み込みニューラルネットワーク(DCNN)

本研究では、AlexNet[1]に基づくネットワークを日本古典文学くずし文字[3]の識別に用いた（図 1）。構築モデルは、畳み込み層(conv)4 層と全結合(liner)3 層から構成される。畳み込み層間を連絡する活性化関数として ReLU を使用した。プーリングは、最大値プーリングを用いた。事前学習は行わず、モデルの初期パラメータはランダムに決定される。モデルの出力と学習データに付与されたラベル間の交差エントロピー誤差を誤差関数として、誤差逆伝播法により学習を行った。最適化手法として Adam[4]を用いた。

2.2 古典文学くずし文字データセットの拡張

DCNN モデルが獲得する識別能力は、学習に用いられるデータ量に依存する。本研究で用いた日本古典文学くずし文字データセットは「おらが春」一冊より抽出された[3]。しかし、データセット内の文字種ごとのデータ数に大きな偏りが確認された。本研究では、データ量が限定される文字種に対して、現代人の模写によるくずし文字データセットの拡張を行った。模写の対象としたくずし文字は、オリジナルデータセット内でデータ量（文字データの個数）が 1 個となる 484 文字種、データ量が 2 個の 189 文字種、データ量が 3 個の 113 文字種、そしてデータ量が 4 個の 57 文字種である。

データセット拡張のためのくずし文字模写は、学生 19 名の実験参加者により行われた。各実験参加者は、筆ペンを用いて、1 つの文字種に対して 4~5 個づつ手書きで模写を行う。くずし文字は、7 センチメートル四方の領域に 1 文字づつ模写された。手書き模写の際、20 インチディスプレイに模写の対象となる文字を実験参加者に呈示し、模写の見本とした。実験参加者により手書き模写されたくずし文字をスキャナで読み取り、トリミングした後、文字種ごとに分類し、データセットを拡張した。実験参加者により模写された手書きくずし文字の例を図 2 に示す。

実験参加者により模写されたくずし文字から、汚損等が確認されたものを除き、1 つの文字種に対して 4 個のくずし文字を拡張データセットに適用した。構築した拡張くずし文字データセットでは、843 文字種のデータ量が 4 倍となった。この拡張されたデータセットを図 1 の DCNN ネットワークへと適用した。本研究では、現代人により模写されたくずし文字を学習データセットして用いた。モデルはオリジナルの日本古典文学くずし文字データセットに含まれるくずし文字を識別する。

3. 結果

前章にて示した DCNN へと拡張したデータセットを適用し、日本古典文学である「おらが春」内に記載されたくずし文字を識別するモデルを獲得させた。ここでは、オリジナルのデータセットにおいてデータ量が少ない文字種を識別対象とした。より高い識別能力を DCNN に獲得させるために、拡張したデータセットを増幅させた（図 3）。増幅量は、拡張したデータセット内のデータ数を 4 倍、10 倍、25 倍、そして 50 倍とした。これらにより、現代人が模写したくずし文字から古典文学特有のくずし文字の識別が可能となるかを検証した。

獲得した DCNN モデルによるくずし文字識別率を増幅率ごとに図 4 に示す。オリジナルデータセット[3]では、データ量が少なく、識別が困難であった文字種に対して、拡張したデータセットから学習したモデルはある程度の正答率を示した。また、学習に用いるデータ量の増幅率を大きくするに伴い、モデルのくずし文字識別率が上昇した。しかし、増幅率 25 倍と 50 倍では、識別率に顕著な差は確認できなかった。

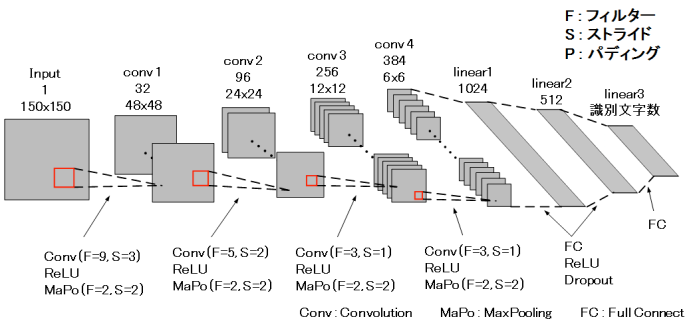

図 1 構築した DCNN モデル

† 東邦大学理学部情報科学科, Toho University
‡ 東京電機大学理工学部情報システムデザイン学系, Tokyo Denki University

4. おわりに

本研究では、日本古典文学に記載されたくずし文字を識別するモデルを獲得するため、くずし文字データセットの拡張を試みた。現代人が手書き模写する事で、古典文学で用いられるくずし文字を拡張させた。拡張したデータセットを DCNN へと適用する事により、オリジナルデータセットではデータ量が限定される文字種に対する識別が可能になった。これは、現代人が手書き模写したくずし文字が、学習用データセットとして利用できる可能性を示唆している。

本研究では、くずし文字データセットを模写する事で拡張した。しかし、その拡張量と拡張した文字種が限られていたため、モデルは充分な識別能力を獲得できなかった。これは、本研究で拡張したデータセットでは、「おらが春」に記載されている文字種の識別に必要となる特徴量を表現しきれなかった可能性が考えられる。また、敵対性生成ネットワーク[5]などを用いる事でより高精度の識別を行うモデルが獲得されると期待される。

謝辞

ご討論頂いた東京電機大学日高章理准教授に感謝する。本研究は科研費(no. 17K12704)の助成を受けたものである。

参考文献

[1] Krizhevsky Alex, Sutskever Ilya, Hinton E. Geoffrey, “Image Net Classification with Deep Convolutional Neural Network”, Proceedings of the 25th International Conference on Neural Information Processing System, Vol. 1, 1097-1105 (2012).
[2] Kerenidis Iordanis, Luongo Alessandro, “Quantum Classification of the MNIST Dataset via Slow Feature Analysis”, arXiv:1805.08837v1 (2018).
[3] 人文学オープンデータ共同利用センター、http://codh.rois.ac.jp/
[4] Kingma P. Diederik, Ba Jimmy, “Adam: A Method for Stochastic Optimization”, arXiv:1412.6980v9 (2017).
[5] Goodfellow Ian, Pouget-Abadie Jean, Mirza Mehdi, Xu Bing, Warde-Farley David, Ozair Sherjil, Courville Aaron, Bengio Yoshua, “Generative Adversarial Networks”, arXiv:1406.2661v1 (2014).

図 2 手書き模写により拡張したくずし文字データセット例

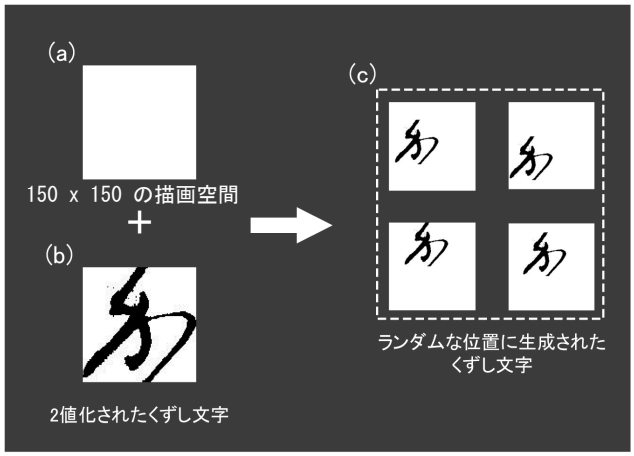

図 3 本研究にて用いたデータ増幅法

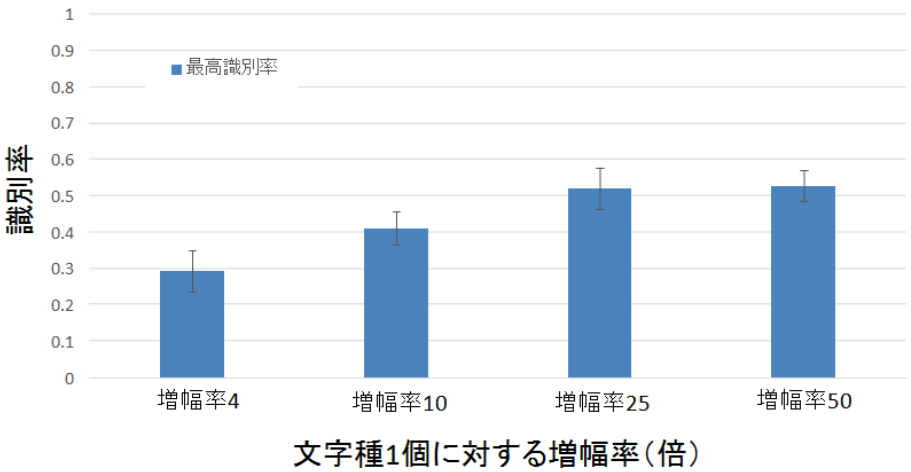

図 4 拡張データセットを用いたくずし文字識別率