CE-CLCNN: Character Encoderを用いたCharacter-level Convolutional Neural Networksによるテキスト分類

北田 俊輔　　彌冨 仁

法政大学 理工学部 応用情報工学科
{shunsuke.kitada.8y@stu., iyatomi@}hosei.ac.jp

概 要

日本語などの単語ごとに明確な区切りがない言語の解析には，一般的に困難な形態素解析を実施した後，それらに対する適切な埋め込みが必要である．また，過学習を抑えるための data augmentation を自然言語処理に適用する場合，意味の解析を要するため通常簡単ではない．本研究ではこれらの問題を低減させ，文書分類を行える end-to-end モデルである character encoder character-level convolutional neural networks (CE-CLCNN) を提案する．CE-CLCNN は，解析する文の各文字を画像として扱うことで，文字の形態に着目した優れた埋め込みを実現するだけでなく，画像認識分野の data augmentation が適用可能となる．また，CNN の持つ卓越した学習能力を文書解析に活かせるため，優良な文書解析能力が実現できる．本報告では，CE-CLCNN が公開されているデータセットに対して state-of-the-art の認識精度を実現した．加えて本稿では CE-CLCNN が文書分類を行う際，解析対象のどの部分に着目しているかについても可視化を行って考察した．

1 はじめに

過学習は機械学習における課題の 1 つであり，特に deep neural networks においては汎化性能を向上させるために様々な正則化手法が提案されている [1, 2]．またデータセットを擬似的に増やす data augmentation も汎化性能を向上させる重要な手段として幅広く適用され，効果が確認されている．自然言語処理における data augmentation は助詞置換や時制変換，シソーラスを用いた単語変換，スペルミスによる単語表現の変化などが挙げられるが，これらは正確な分かち書きが必要であり，また効果的な拡張のためには意味の解釈が必要であることから総じて簡単ではない．

ここで英語と比較して日本語や中国語，韓国語といったアジア圏の言語では，一般的に正確な分かち書きが難しいとされている．この問題に対して近年では単語分割が不要な character-level の特徴を用いた文書分類手法が提案されてきた [3, 4]．Character-level convolutional neural networks (CLCNN) [3] は画像認識分野で素晴らしい成果を上げている convolutional neural networks (CNN) の入力を 1 次元にしたもので，one-hot 表現とした各文字や，lookup table 等の埋め込み手法を適用させて得られたベクタを入力とし，一次元方向に畳み込むことにより学習，識別を行う．このため CLCNN は入力する文書の分かち書きが不要で，なおかつ識別などに必要な特徴を自動で学習できる強力な解析手法である．しかしながら日本語や中国語など，文字種の多い言語では過学習を引き起こしてしまう問題がある．

この問題に対して文字列を構成する各文字を画像とみなし、前段で文字画像を入力とした convolutional auto-encoder (CAE) [5] や PCA による低次元の文字の埋め込み表現を学習させた後，その文字表現を用いて後段の CLCNN の学習を行うモデルが，島田らによって提案されている [4]．この手法は文字を画像として扱うことにより，各文字の形状を活かした文字のベクタ表現の獲得が期待できる．CNN をベースとした CLCNN では，汎化性能を高めるために大量のテキストデータが必要であることが知られており，データ数が不足している場合には，data augmentation によって擬似的にデータ数を増やすことが必要である．彼らはこの問題に対して形態素解析が不要な data augmentation として，wildcard training (WT) を提案している．これは学習時の入力文書の一部である文字ベクトルの任意の要素に対して，ランダムに dropout [1] させることで汎化性能の大幅な向上を狙った手法で，青空文庫の著者推定や Web 新聞記事における新聞社推定等のタスクにおいて，精度 10%程度向上を確認している．一方，このモデルは CAE と CLCNN の学習を別に行っているために，性能向上の余地が残されていた．また文字を画像として扱っているメリットを十分に活かしきれておらず，さらなる改善の余地が見込まれた．

以降 Liu ら [6] は，CNN を用いた character encoder

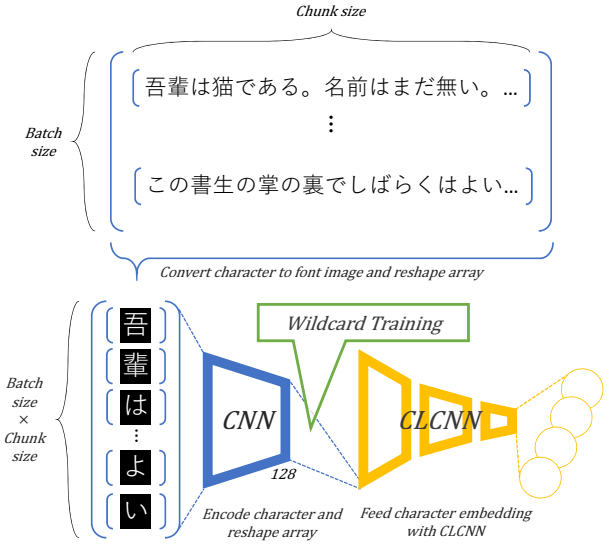

図 1: 提案手法 (CE-CLCNN) の全体像

と gated recurrent unit (GRU) [7] を用いて，日本語や中国語，韓国語の文字画像から文字表現の獲得と文書分類まで end-to-end で学習するモデルを提案している．このモデルでは，島田らのモデルと異なり明示的に文字形状を保持するように学習を行っていないが，形状的特徴が似ている文字が近い文字表現になっていることが示されている．また，中国語の文字画像に対して CAE を適用し，その文字表現を用いて単語表現を獲得する Su らの手法 [8] も，学習された文字表現において，形状的特徴が似ている文字に対して近い文字表現を獲得できていることが示されている．

こうした背景から本研究では，解析する文の各文字を画像として，文字形状に応じた低次元の埋め込み表現を獲得し，文字表現の学習から文書分類まで一括して行える end-to-end モデルである，character encoder character-level convolutional neural networks (CE-CLCNN) を提案する．CE-CLCNN は，これまでの日本語などの言語の解析で問題であった以下の困難さ: (1) 形態素解析の困難さ，(2) 文字種の多様性から引き起こされる過学習，(3) 自然言語処理における data augmentation の困難さ を解決し，高い文字解析能力と汎化性を備えたモデルである．

2 CE-CLCNN

提案する CE-CLCNN の全体像を図 1 に示す．CE-CLCNN は文字画像から文字表現を学習する character encoder (CE) と文書分類を行う CLCNN で構成させている．これらのネットワークを構成する畳み込み

表 1: Character Encoder (CE) のアーキテクチャ

Layer #	CE configuration
1	Conv(k=(3, 3), o=32) → ReLU
2	Maxpool(k=(2, 2))
3	Conv(k=(3, 3), o=32) → ReLU
4	Maxpool(k=(2, 2))
5	Conv(k=(3, 3), o=32) → ReLU
6	Linear(800, 128) → ReLU
7	Linear(128, 128) → ReLU

表 2: CLCNN のアーキテクチャ

Layer #	CLCNN configuration
1	Conv(k=(1, 3), o=512, s=3) → ReLU
2	Conv(k=(1, 3), o=512, s=3) → ReLU
3	Conv(k=(1, 3), o=512) → ReLU
4	Conv(k=(1, 3), o=512)
5	Linear(5120, 1024)
6	Linear(1024, # classes)

カーネルなどのパラメータはクロスエントロピー誤差関数を目的関数とし，誤差逆伝播法によって最適化される．

2.1 Character encoder(CE) による文字表現の獲得

CE では文書の各文字を 36 × 36 サイズの文字画像に変換したうえで，CE を用いて各文字画像から d_{CE}bit の文字表現を学習する．本実験でのアーキテクチャでは，CE は CNN の構成を取っており，文書から文字列長 $\mathcal{C}$ のチャンクを $\mathcal{B}$ 個ずつ順次入力として受け取る．つまりバッチサイズが $\mathcal{C} \times \mathcal{B}$ となるように reshape し，CE への入力としている．今回使用した CE のアーキテクチャを表 1 に示す．ここでカーネルサイズ k，フィルター数 o である．

2.2 Character-level convolutional neural networks (CLCNN) による文書分類

CE からエンコードされた d_{CE}bit/文字の文字表現に対して再度文字列長 $\mathcal{C}$ でバッチサイズ $\mathcal{B}$ になるよう reshape し，CLCNN の入力とする．入力に対して畳み込み処理を行う際，自然言語処理で広く用いられるプーリング処理の代わりにスライドサイズ s を調整

図 2: Random erasing data augmentation [9] を適用した文字画像の例

表 3: Random erasing のパラメータ

Parameter	Scale
Erasing probability p	0.3
Max area ratio s_l	0.4
Min area ratio s_h	0.02
Max aspect ratio r_1	2.0
Min aspect ratio r_2	0.3

したアーキテクチャを用いた．今回使用した CLCNN のアーキテクチャを表 2 に示す．

2.3 Random erasing と Wildcard training による Data augmentation

CE に入力される文字画像に対して random erasing data augmentation (RE) [9] を適用する．各文字画像にはランダムに矩形領域をノイズでマスクし，図 2 のように文字の一部分が隠されているような学習データを生成する．Random erasing data augmentation で用いる各パラメータを表 3 に示す．

CE でエンコードされた文字表現に対して wildcard training (WT) [4] を適用する．Wildcard training を適用した文字表現を CLCNN に入力することで，wildcard となった文字についての情報が CLCNN に伝搬しないため，wildcard となった文字以外の文字間共起から，正しい推定をしなければならなくなる．よって，wildcard training によって学習された CLCNN はより汎化性能が高くなることが期待される．

3 実験

3.1 実験設定

文字の埋め込み表現の次元数とチャンクサイズをそれぞれ $d_{CE} = 128$ と $\mathcal{C} = 128$ とした．wildcard training における wildcard ratio は $\gamma_w = 0.1$ に設定した．バッチサイズは $\mathcal{B} = 256$ とし，パラメータの最適化には Adam [10] を使用した．

モデルの学習時には文書中から連続する文字列長 $\mathcal{C}$ 分だけチャンクとして取り出して学習に使用し，評価時にはチャンク $\mathcal{C}$ をスライドサイズ $s = 1$ でスライドさせ，文書全体を入力として使用し，評価を行った．

表 4: Wikipedia タイトルのカテゴリ推定の結果

Method	Accuracy[%]
(Proposed) RE + CE-CLCNN + WT	**58.4**
(Proposed) RE + CE-CLCNN	58.0
(Proposed) CE-CLCNN	54.4
CLCNN + WT [4]	54.7
CLCNN [4]	36.2
VISUAL model [6]	47.8
LOOKUP model [6]	49.1
Ensemble (VISUAL + LOOKUP) [6]	50.3

3.2 Wikipedia タイトルのカテゴリ推定

Wikipedia タイトルのカテゴリ推定タスクでは，[6] で用いられている，Wikipedia から記事タイトルを収集したデータセットのうち，日本語のものを利用した．

Geography, Sports, Arts 等の 12 クラス，計 206,313 タイトルから 8 割を学習，2 割を評価とする実験を行った．前処理として，入力される文書が 10 文字以上になるようゼロパディングを行った．

これらのデータセットを用いた評価実験の結果を表 4 に示す．提案する CE-CLCNN が先行研究の手法を上回る結果となった．島田ら [4] の結果は，それ以降に提案された Liu らの手法 [6] よりも 4%程度良好な結果であったが，提案手法はさらに 4%程度優れた結果となり，著者の知りうる限りこの公開問題に対して最良の結果を達成した．特に画像に対する拡張である RE と文字表現に対する拡張である WT を併せて適用することで，より汎化性能の高いモデルになっていることが確認できた．

3.3 Character Encoder の分析

Wikipedia タイトルデータセットを使用して学習させた character encoder を用いて得た文字表現に対して，5-nearest neighbor を適用した結果を表 5 に示す．クエリ文字から得られた近傍文字は部首が似ているものが多く，character encoder が文字の形状的特徴を捉えて学習していることが確認できた．

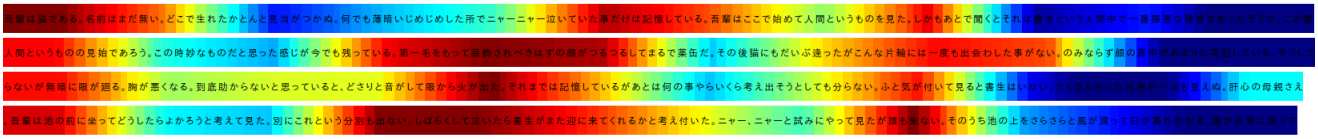

図 3: GradCam を用いたモデルの予測の可視化

表 5: クエリ文字に対する文字表現の近傍上位 5 文字

クエリ文字	近傍文字	ユークリッド距離
鮫	鰭	370.1
	駮	403.7
	鮪	405.2
	鰐	409.4
	鰤	409.6
痛	癨	317.2
	癪	388.3
	癒	398.3
	痕	398.9
	痴	399.2
披	彼	452.8
	擅	491.5
	擔	520.5
	擒	533.8
	捗	536.8

3.4 Grad-CAM を用いたモデルの可視化

CNN の入力に対してモデルが解析対象のどの部分に注目して予測を行っているかを可視化することができる Grad-CAM [11] を，可視化の効果が考察しやすいと考えられる青空文庫データセットで学習した CE-CLCNN に対して適用した．青空文庫[1]で公開されている文書のうち，対象著者 10 名の作品 10 編ずつ計 104 作品から本文のみを抽出したデータセットを利用し，10 クラスの著者推定問題として本モデルを学習させた．データの 8 割を学習，2 割を評価とする実験において提案する CE-CLCNN は 73.9%の精度を実現した．この数値は島田ら [4](69.6%) を上回るもので，CE-CLCNN が文学作品から著者を推定する手がかりを獲得できていることが確認できた．その後学習が終わったモデルに夏目漱石の作品の *吾輩は猫である* の先頭 512 文字を入力とし，各チャンクに対して，Grad-CAM を適用した結果を図 3 に示す．

この結果は 1 例でしかないが，*吾輩は猫である* で主人公を表す特徴的な一人称 *吾輩* 付近が活性化しており，モデルが特徴的な言葉に対して反応していることが確認できた．

[1] 青空文庫 Aozora Bunko http://www.aozora.gr.jp/

4 おわりに

本研究では end-to-end で文字表現の獲得から文書分類を行う CE-CLCNN を提案した．文字画像を CE でエンコードして得られた文字表現は形状的特徴を捉えていることを確認し，それらの文字表現を入力とした CLCNN によって，単語分割不要で高精度な日本語の文書分類を可能にした．画像認識分野の data augmentation と既存の強力な正則化手法を組み合わせることにより，さらなる汎化性能の向上を示した．さらに，モデルの可視化手法を用いることでモデルの解釈性を高めることができた．

参考文献

[1] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR arXiv:1207.0580*, 2012.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[3] X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” *Advances in neural information processing systems*, pp. 649–657, 2015.

[4] D. Shimada, R. Kotani, and H. Iyatomi, “Document classification through image-based character embedding and wildcard training,” *IEEE International Conference on Big Data*, pp. 3922–3927, 2016.

[5] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 52–59, 2011.

[6] F. Liu, H. Lu, C. Lo, and G. Neubig, “Learning character-level compositionality with visual features,” *The 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *CoRR arXiv:1412.3555*, 2014.

[8] T. Su and H. Lee, “Learning chinese word representations from glyphs of characters,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pp. 264–273, 2017.

[9] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” *CoRR arXiv:1708.04896*, 2017.

[10] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR arXiv:1412.6980*, 2014.

[11] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization,” *CoRR arXiv:1610.02391*, 2016.