文字の構成要素を考慮した文書分類の再考

津嶋, 祐介 / TSUSHIMA, Yusuke

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学研究科編
(巻 / Volume)
65
(開始ページ / Start Page)
1
(終了ページ / End Page)
5
(発行年 / Year)
2024-03-24
(URL)
https://doi.org/10.15002/00030748

文字の構成要素を考慮した文書分類の再考

RETHINKING DOCUMENT CLASSIFICATION CONSIDERING CHARACTER COMPONENTS

津嶋 祐介
Yusuke TSUSHIMA
指導教員 彌冨仁

法政大学大学院理工学研究科応用情報工学専攻修士課程

In Japanese and Chinese natural language processing, character-based natural language processing that takes into account the radicals of Chinese characters contributes to the improvement of document analysis performance.In order to take character shapes into account, many end-to-end models of character encoders and document classifiers based on convolutional neural networks (CNNs) have been proposed in the past. In this study, we propose Vision Transformer - Character-level Transformer (ViT-CLT), which consists of a Vision Transformer (ViT) as a character encoder and a character-level Transformer (CLT) as a document classifier, in order to realize high-performance document classification that takes into account relationships among character sub-components. In evaluation experiments, we confirmed that ViT-CLT improved prediction performance by 20.0% compared to previous character encoder and document classifier models using CNNs in a categorization task using Japanese news articles.Furthermore, the visualization results of attention in ViT-CLT's character encoder showed that ViT-CLT was able to consider the components of Chinese characters better than the previous model.

Keywords *: natural language processing, character embedding, deep learning*

1. はじめに

日本語や中国語といったアジア圏の言語に対する自然言語処理において，その特徴的な文字形状を考慮することが文書解析能力の向上に繋がることが知られている．日本語や中国語で主に使用されている漢字の字体において，その部分をなす点画の一定のまとまりは “構成要素” と呼ばれ，偏や旁などの部首やその漢字自体の場合もあるが，それらに限定されない点画のまとまりも含まれている．このような文字の構成要素における形状的特徴やその位置関係を捉えることで，自然言語処理モデルの表現力向上が期待できる [1, 2, 3].

文書を文字単位で扱う先行研究の多くは，畳み込みニューラルネットワーク (convolutional neural network; CNN) からなる文字符号化器から文字形状を考慮した埋め込みを学習している．これらの手法は局所的な情報である漢字の偏旁冠脚といった構成要素の単体を考慮することができる．更にこうした埋め込みを利用した文書解析モデルは，高い性能を実現しているしかしながら文字符号化器に CNN を利用したモデルは漢字の各構成要素の形状的特徴を捉えることに成功する一方で，文字構成要素間の位置関係の考慮に改良の余地が見受けられた．

入力情報の局所的な特徴に加えて大局的な特徴の関係性を考慮できる Transformer [4] を画像認識タスクに応用した Vision Transformer (ViT) [5] は，一般的な画像認識タスクで CNN を超える予測性能が報告され始めている．ViT は入力画像を複数のパッチに分割し Transformer に入力することで，CNN よりも大局的な情報を学習可能であり，位置埋め込みを各パッチに付加することで，パッチの位置関係を考慮することができる．さらに self-attention の効果により，パッチ間の関係性を考慮することが可能である．文字画像の各パッチが構成要素の一部または全体の情報を持っている場合，パッチ間の関係性を考慮できれば構成要素間の関係性を考慮できると考えている．こうした性質から，文字画像から構成要素といった局所的形状を捉えるとともに，構成要素間の位置関係を捉えられると考えられるため，ViT を文字符号化器として用いることで文字構造を捉えたより良い文字の埋め込みの獲得が期待できる．

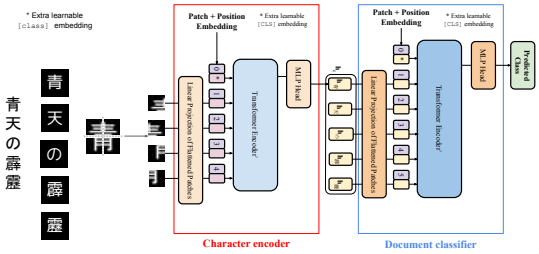


Fig. 1: **提案モデル ViT-CLT の全体図**

本研究では，偏旁冠脚といった文字の構成要素の位置関係を考慮することで高性能な文書分類を目指す ViT-CLT を提案する．提案手法は文字符号化器に ViT，文書

分類器に Transformer で構成され，これら end-to-end で学習を行う．文字符号化器を担う ViT は，文字内の構成要素に対してその形状的特徴とその関係性を捉えることができる．また文書分類器を担う character-level Transformer (CLT) は，近傍の文字以外との関係性も考慮できるため，大局的な文脈を考慮できる．

2. 提案モデル

我々は，漢字の偏旁冠脚といった構成要素とその関係性を考慮した文書分類モデルである ViT-CLT を提案する．提案する ViT-CLT の全体像を図 1 に示す．文字を文字画像に変換し，ViT からなる文字符号化器で文字埋め込みを得た後，それらを元に Transformer からなる文書分類器で文書分類を end-to-end で学習する．

(1) ViT を元にした文字符号化器

ViT-CLT の文字符号化器は ViT を用いることで，各文字画像の局所構造の関係性を考慮した符号化を実現する．ViT [5] は Transformer [4] を基にした画像処理モデルである．入力画像をいくつかの部分的な画像 (パッチ画像) に分割し，画像の局所的な情報を取得する．ViT が従来の CNN より優れている点は，各パッチ画像に位置エンコーディングを付加し，Transformer に入力することで各パッチ画像間の関係性を捉えることができる点である．入力を文字画像にすることで，文字の構成要素の形状的特徴内の局所的な情報と，構成要素間の大局的な関係性を捉えることが期待できる．入力画像の埋め込みとして，[CLS] トークンに相当する ViT の出力を用いた．

(2) Transformer を元にした文書分類器

ViT を元にした文字符号化器から得られた文字埋め込みを元に，character-level Transformer (CLT) を元にした文書分類器を学習する．従来の character-level CNN (CLCNN) は畳み込み処理によって入力テキストの文字同士の局所的な関係性を考慮しつつ文書分類できるよう訓練されていた．本研究で用いる CLT は各文字に対して self-attention を計算することで CLCNN に比べてさらに離れた文字同士の関係性も考慮可能である．これにより，提案する ViT-CLT は各文字内の構成要素という，これまでより細かい局所的な特徴から，文脈中の文字の関係性という大局的な特徴まで捉えることができる．

3. 実験

提案モデル ViT-CLT は ViT を元にした文字符号化器と Transformer を元にした分類器の end-to-end モデルである．このモデルに対して従来モデルである，文字符号化器が CNN，文書分類器が CLCNN の CE-CLCNN [6] と比較した．更に我々は 文字符号化器 を CNN と ViT， 文書分類器を CLCNN と CLT に変えたモデルの性能を比較した．

比較実験として，以下のような実験を行った．

- **文字符号化器を CNN と ViT で比較した際の CLT の文書単位の attention の可視化**: attention の当たり方から，文書分類の際にどの文字に着目しているかを考察した．
- **パッチ分割手法を従来の分割手法と SWP で比較した際の文字単位の attention の可視化**: attention の当たり方から，文字符号化の際にどのパッチに着目しているかを考察した．
- **対象文字埋め込みの近傍の文字**: 文字の近傍にどのような文字が埋め込まれているのかを調べた．CNN や ViT によって文字符号化した埋め込み表現はどのような性質を持っているのかを考察した．

文字符号化器は MAE を用いて事前学習を行った。学習データは常用漢字、ひらがな、アルファベットなどを含む 6632 文字、入力には 60 × 60 のグレースケール画像を使用した。埋め込み次元は CNN と ViT でそれぞれ 128 次元に設定した．モデルは精度が最も良かった分割数のものを使用した．モデルの最適化には Adam [7] を用いて，バッチサイズ 64, エポック数 100 で訓練した．また，attention 可視化方法は，rollout [8] という attention スコア同士の行列積の結果を利用する手法を使用した．

実験用データセットとして livedoor ニュースコーパスと Multilingual Amazon Reviews Corpus を使用し，記事のカテゴリを分類するタスクとレビューの極性を分類するタスクで提案法を評価した．

(1) livedoor ニュースコーパス

日本語ニュース記事データセットとして livedoor ニュースコーパス[*1] を使用した．このデータセットには計 9 カテゴリの記事が含まれているカテゴリ分類時には記事のタイトルからそのカテゴリを分類できるようにモデルを学習した．データセットの 8 割を学習用，2 割を評価用に各クラスが均等になるように分割を行った．前処理として文字列長さは 128 になるように先頭から 128 文字を切り出した．

(2) Multilingual Amazon Reviews Corpus

感情分類データセットとして Multilingual Amazon Reviews Corpus [9] を使用した。

このデータセットには 2015 年から 2019 年の間に収集された、英語、日本語、ドイツ語、フランス語、スペイン語、中国語のレビューが含まれており、データセット内の各レコードには、テキスト、タイトル、評価 (1〜5)、匿名化されたレビュー者の ID、匿名化された製品 ID、および大まかな製品カテゴリが含まれている。我々は日本語のレビューの本文から評価の極性を推定するように、モデルを学習した。スター数が 1〜2 のレビューはネガティブ、スター数が 4〜5 のレビューはポジティブとラベリングを行ったものを学習データとして使用した。データセットの 8 割を学習用，2 割を評価用に各クラスが均等になるように分割を行った．前処理として文字列長さは 256 になるように先頭から 256 文字を切り

[*1] http://www.rondhuit.com/download.html\#ldcc

出した.

4. 結果・考察

実験結果では CLCNN と CLTransformer によるカテゴリ分類について示す．また，パッチを分割した時の精度の変化や文章内の attention の当たり方についても述べる．

(1) livedoor ニュースコーパスによるカテゴリ分類

表 1 にニュース記事のカテゴリ分類性能の比較結果を示す．提案モデル ViT-CLT が従来モデルである CE-CLCNN [6] の macro F1-score を 20.0% 超える大幅な予測性能向上を確認した．従来モデルの文字符号化器を ViT にした場合（ViT + CLCNN）は 8.7% の向上が確認された．更に従来モデルの文書分類器を Transformer にした場合 (CNN + CLT) は 14.2% の向上が確認された．

Table 1: **livedoor ニュースコーパスによるカテゴリ分類の結果**

モデル	文字符号化器	文書分類器	F1.
CE-CLCNN [6]	CNN	CLCNN	0.615
CE-CLCNN	ViT	CLCNN	0.708
&	ViT + SWP	CLCNN	0.659
ViT-CLT variants	CNN	CLT	0.757
ViT-CLT	ViT	CLT	**0.815**
(提案法)	ViT + SWP	CLT	0.785

よって、文書分類器に Transformer からなるモデルを採用することで予測性能の向上に大きく寄与することが確認された．また，提案モデルにおいて，従来のパッチ分割を採用した際にパッチ分割数を 9 に設定した場合が最も良い性能となった．しかし，我々の SWP を適用した場合は一定の予測性能の向上に寄与した一方で，SWP を適用しない場合に少し劣る結果となった．SWP は類似した特徴を持つパッチ画像同士の関係を学習するため，文字の構成要素の構造を大きく捉えてしまったことにより，構成要素間の関係を捉えることが難しくなったことが影響したと考えられる．同様の傾向が ViT + CLCNN の場合にも確認された．

(2) Multilingual Amazon Reviews Corpus によるレビューの極性分類

表 2 にニュース記事のカテゴリ分類性能の比較結果を示す．提案モデル ViT-CLT が従来モデルである CE-CLCNN の macro F1-score を 2.8% 超える大幅な予測性能向上を確認した．従来モデルの文字符号化器を ViT にした場合（ViT + CLCNN）は 3.0% の向上が確認された．しかし、従来モデルの文書分類器を Transformer にした場合 (CNN + CLT) は 0.2% 下がった。

提案モデルにおいて，SWP を採用した際にパッチ分割数を 9 に設定した場合が最も良い性能となった．よって、文字符号化器に ViT からなるモデルを採用することで予測性能の向上に大きく寄与することが確認された。また、提案モデルにおいて、従来のパッチ分割を採用した際にパッチ分割数を 9 に設定した場合が最も良い性能となった。レビューの極性分類の予測性能の向上に文書解析器よりも文字符号化器の寄与が大きかったことから、我々の手法による文字埋め込みが以前の手法よりもより適切に意味を捉えられていると考えられる。

Table 2: **Multilingual Amazon Reviews Corpus によるレビューの極性分類の結果**

モデル	文字符号化器	文書分類器	F1.
CE-CLCNN [6]	CNN	CLCNN	0.812
CE-CLCNN	ViT	CLCNN	0.838
&	ViT + SWP	CLCNN	0.829
ViT-CLT variants	CNN	CLT	0.810
ViT-CLT	ViT	CLT	0.839
(提案法)	ViT + SWP	CLT	**0.840**

(3) 文書単位の attention の可視化

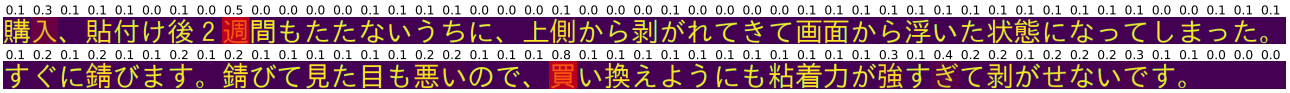
図 2 の結果から CLT は予測に重要な情報 (e.g., 単語粒度の情報) をより適切に捉えらえていたため, CLCNN よりも高い精度が得られたと考えられる．CLT は各文字に位置エンコーディングの付与と attention により，文書の近傍以外の文字間の関係と予測に重要な情報を捉えているため，CLCNN よりも高い精度が得られたと考えられる．さらに，提案モデルの文字符号化器である ViT を使用した場合はより単語粒度の情報に着目していたため，ViT の文字符号化器は文字の意味を捉えた文字符号化を行っていると考えられる．さらに、SWP を適用した場合、「優」や「悪」といった意味のある文字に対してより着目していたことから、従来のパッチ分割手法よりも文字の意味を捉えた文字符号化を行っていると考えられる．

(4) 文字単位の attention の可視化

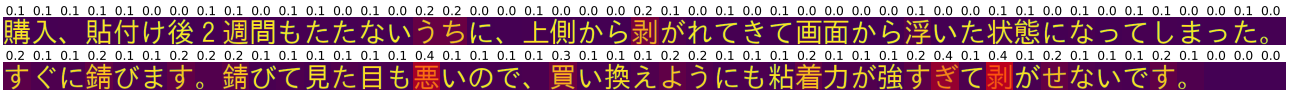
図 3 の結果から文字の形状を捉えた attention が当たっており，漢字の場合は部首を捉えている．SWP の attention の可視化から，従来のパッチ分割と同様に文字の形状を捉えた attention が当たっている．さらに SWP では漢字の部首だけでなく，部首以外の部分にも attention が当たっていることから，構成要素間の構造を捉えられていると考えられる．これは SWP を用いることで局所的な情報を捉えやすくなったためだと考えられる．しかし，似た文字領域に attention が当たってしまっている場合がある．これは同じようなパッチが複数存在するため， attention が分散しやすくなってしまったと考えられる．

(5) 対象文字埋め込みの近傍の文字

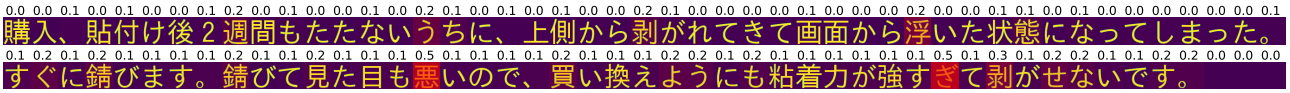
表 3 に各比較モデルにおける “語” の埋め込みの近傍 5 字を示す．文字符号化器に CNN を使用した場合と従来のパッチ分割を使用した ViT を使用した場合で比較すると，ViT の方が偏旁冠脚といった構成要素をより捉えられていた．一方従来のパッチ分割を使用した場合と SWP を使用した場合で比較すると，SWP は偏旁冠脚といった構成要素を捉えられなかった．SWP が偏旁冠脚といった構成要素を捉えられなかったのは，attention



(a) Character encoder: CNN

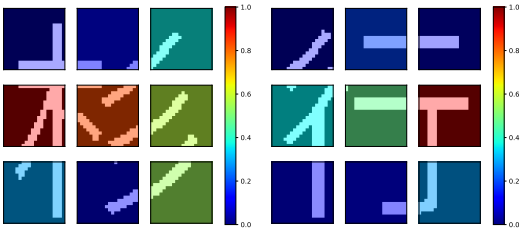


(b) Character encoder: ViT (**提案手法**)

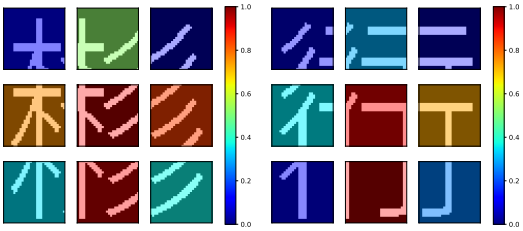


(c) Character encoder: ViT + SWP (**提案手法**)

Fig. 2: **文字符号化器を CNN と ViT で比較した際の CLT の文書単位の attention の可視化**: 上部の数字は実際の attention のスコアを示す.



(a) 従来のパッチ分割 (パッチ数 = 9)



(b) SWP (パッチ数 = 9)

Fig. 3: **提案モデルの文字単位におけるパッチ間の attention の可視化**

Table 3: **漢字 “語” の近傍の漢字上位 5 つ** : 括弧内の数字は分割数を示す. 通常のパッチ分割はほかの文字符号化器よりも部首を捉えられている.

文字符号化器	最近傍		···		5 番目
CNN	訟	諮	諂	讎	誚
ViT (16)	牾	調	諤	琣	銀
ViT + SWP (9)	認	返	議	謂	託

が分散して，局所的な構造単体を捉えることが難しくなってしまったと考えられる.

5. おわりに

日本語や中国語において，漢字を中心に文字の形状を考慮する自然言語処理モデルが提案されてきたが，従来手法では偏旁冠脚といった文字の構成要素間の関係を考慮することは難しかった．本研究では，文字の構成要素間の関係を捉えることができる ViT-CLT を提案した．ViT からなる文字符号化器で偏旁冠脚個別の形状を捉えるとともに，構成要素間の関係性も捉えることを可能とした．日本語のニュース記事のカテゴリ分類による評価実験の結果，我々の提案手法が従来の CNN + CLCNN モデルを遥かに超える性能を実現した．更に提案手法の文単位および文字単位の attention 可視化を通じて，我々の手法が従来手法よりも構成要素間の関係性を適切に捉えていること確認した.

今後は Vision Transformer を文字の形状的特徴をより効果的に考慮できるよう改良することを検討している。

謝辞

本研究にあたり，全般にわたるご指導をしてくださった彌冨仁教授，共に研究における実験や検討を重ねた知的情報処理研究室の皆様に深く御礼申し上げます．また，生活面で様々なサポートをしてくださった家族に感謝申し上げます.

参考文献

[1] Y. Sun, L. Lin, N. Yang, Z. Ji, and X. Wang, “Radical-enhanced chinese character embedding,” in *Neural Information Processing* (C. K. Loo, K. S. Yap, K. W. Wong, A. Teoh, and K. Huang, eds.), (Cham), pp. 279–286, Springer International Publishing, 2014.

[2] X. Shi, J. Zhai, X. Yang, Z. Xie, and C. Liu, “Radical embedding: Delving deeper to Chinese radicals,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, (Beijing, China), pp. 594–598, Association for Computational Linguistics, July 2015.

[3] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo, “Styleswin:

Transformer-based gan for high-resolution image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11304–11314, June 2022.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[6] S. KITADA, R. KOTANI, and H. IYATOMI, "End-to-end text classification via image-based embedding using character-level networks," in *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–4, 2018.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[8] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 4190–4197, Association for Computational Linguistics, July 2020.

[9] P. Keung, Y. Lu, G. Szarvas, and N. A. Smith, "The multilingual amazon reviews corpus," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.