ViT-CLT: パッチ分割した文字画像から偏旁冠脚を考慮した文書分類

津嶋 祐介 [1]　青木 匠 [2]　北田 俊輔 [2]　彌冨 仁 [1,2]
[1] 法政大学 理工学部 応用情報工学科
[2] 法政大学 理工学研究科 応用情報工学専攻
{yusuke.tsushima.4z, takumi.aoki.4g, shunsuke.kitada.8y}@stu.hosei.ac.jp
iyatomi@hosei.ac.jp

概要

日本語や中国語における自然言語処理では，漢字の部首を考慮した文字単位による自然言語処理が文書解析能力の向上に寄与している．文字形状を考慮するために，従来は convolutional neural network (CNN) を元にした文字符号化器および文書分類器の end-to-end モデルが多く提案されている．本研究では，漢字の偏や旁といった構成要素の関係性を考慮した高性能な文書分類を実現するために，文字符号化器に Vision Transformer (ViT)，文書分類器に character-level Transformer (CLT) で構成された ViT-CLT を提案する．我々の ViT-CLT は漢字の構成要素とその関係性を捉えるために ViT を用いて文字画像から文字の埋め込みを獲得し，文書分類器ではその文字埋め込みを使用して文書分類タスクを解けるよう学習を行う．評価実験では日本語のニュース記事を用いたカテゴリ分類タスクにおいて，CNN を用いた従来の文字符号化器・文書分類器モデルと比較し，ViT-CLT が 18% の予測性能の向上を確認した．更に ViT-CLT の文字符号化器における attention の可視化結果から，従来モデルよりも漢字の構成要素を十分に考慮できていることを確認した．

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 [1, 2]．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” と呼ばれている [3]．この構成要素は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 [4]．

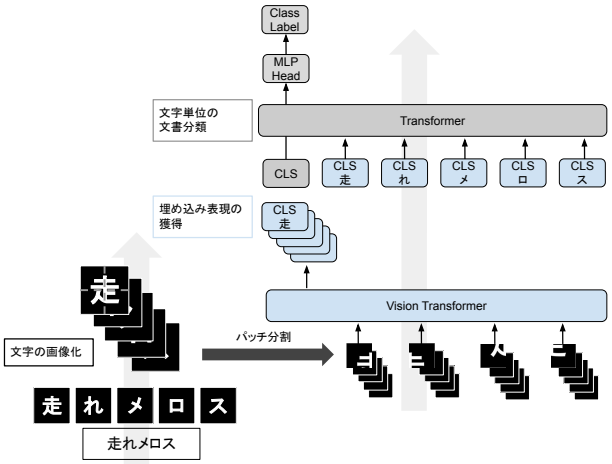


図 1: **提案モデル ViT-CLT の全体図**: 入力文書を文字単位で画像化し，ViT を元にした文字符号化器に入力できるようパッチ分割する．文字符号化器によって得られた文字埋め込みを更に character-level Transformer (CLT) を元にした文書分類器に入力し，[CLS] トークンを元に MLP で文書分類タスクを解く end-to-end モデルである．

文字形状を効果的に考慮するために，文字を画像として深層学習モデルに入力して文字表現を学習する手法が複数提案されている [5, 6, 7, 8, 9]．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 [4, 8]．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 [9].

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 [10] を画像認識タスクに応用した Vision Transformer (ViT) [11] 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．

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．ViT-CLT では多様な文字の構成に対応するため，ViT の画像パッチ切り出しにおいて sliding window algorithm を用いた sliding window patch (SWP) を導入した．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．

評価実験では日本語のニュース記事カテゴリ分類タスクを用いて従来手法比較し，従来手法 [7] である文字符号化器と文書分類器に CNN を利用した時よりも，提案手法は 18%精度向上が確認できた．ViT-CLT の文書単位および文字単位の attention の可視化により，文書単位では，従来の文字単位を入力とするモデルよりも文書解析に影響を与える重要な単語に注意が当たっていることを確認し，さら文字単位では ViT による文字符号化器が文字の構成要素それぞれに注意が当たっていることを確認した．

2 ViT-CLT

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．

2.1 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．漢字は種類が多く多様な構造を持つため，通常 ViT で用いられるパッチ分割手法は多様な文字の構造情報の抽出には不十分であると考えられる．我々は，様々な漢字の構成部分の形状情報を効果的に扱うために，画像パッチの獲得に sliding window patch (SWP) を提案する．SWP は，sliding window algorithm により，各文字から重複を許しながら画像パッチを入力することで，重要な構成要素間の関係の見逃しを防ぐことが期待できる．

2.1.1 Vision Transformer

ViT [11] は Transformer [10] 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．

2.1.2 Sliding Window Patch (SWP)

SWP はパッチ 1 つあたりの特徴を増やすために，sliding window algorithms によりパッチ分割を行う．文字画像に対する従来のパッチ分割では文字画像自体の情報量が少ないため，1 パッチ当たりの情報量がさらに少なくなってしまうという問題の解決を期待する．SWP により局所的な情報をより詳細に得ることができるため，構成要素における部首等の部分構造を学習しやすくなると考えられる．

2.2 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 実験設定

3.1 比較手法

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 [7] 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．

文字符号化器の入力には 128 文字の文書をそれぞれ 60×60 のグレースケール画像に変換して入力した．埋め込み次元は CNN と ViT でそれぞれ 128 次元に設定した．モデルの最適化には Adam [12] を用いて，バッチサイズ 64, エポック数 100 で訓練した．

3.2 データセット

実験用データセットとして日本語ニュース記事コーパスを使用し，記事のカテゴリを分類するタスクで提案法を評価した．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．記事の文字列の長さは，最大 142，最小 6，平均 38 である．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訓練した．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．

1） http://www.rondhuit.com/download.html#ldcc

表 1: ニュース記事のカテゴリ分類の結果

モデル	文字符号化器（パッチ数）	文書分類器	Acc.
CE-CLCNN [7]	CNN	CLCNN	0.604
CE-CLCNN & ViT-CLT variants	ViT (1)	CLCNN	0.653
	ViT + SWP (9)	CLCNN	0.619
	CNN	CLT	0.752
ViT-CLT（**提案法**）	ViT (16)	CLT	**0.785**
	ViT + SWP (25)	CLT	0.781

4 実験結果と議論

提案する ViT-CLT と従来モデルを中心に実験した結果について議論する．更に提案手法の文字単位および文書単位の attention 可視化を通じて，提案モデルが従来モデルよりも偏旁冠脚といった構成要素間の関係性を適切に捉えていることを確認した．

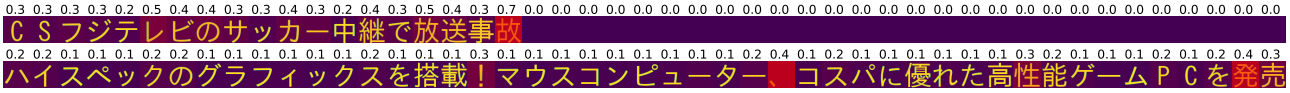
4.1 カテゴリ分類の性能比較

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 [7] の精度 を 18% 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 5% 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 15% の向上が確認された．この結果から文書分類器に 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．

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 16 に設定した場合が最も良い性能となった．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．

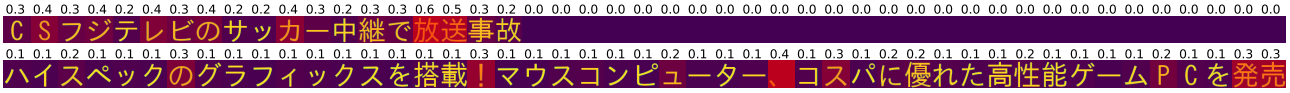
4.2 Attention の可視化

文書単位における文字間の attention 図 2 は文字符号化器に CNN と ViT を使用した場合の CLT の文書単位の attention の可視化結果である．CLT の



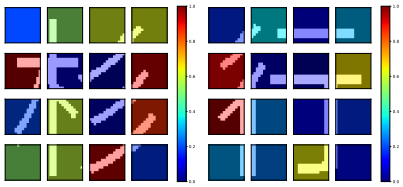
(a) Character encoder: CNN

(b) Character encoder: ViT (**提案手法**)

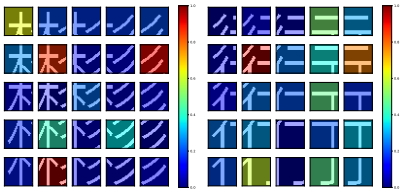


(c) Character encoder: ViT + SWP (**提案手法**)

図 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．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．



(a) 従来のパッチ分割 (パッチ数 = 16)



(b) SWP (パッチ数 = 25)

図 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**: 従来のパッチ分割 (3a) では “きへん”，“ぎょうにんべん” といった部首のみを捉えられている一方で，提案法の SWP (3b) では部首部分と非部首部分双方に attention が当たっている．

attention 可視化方法は，rollout [13] 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．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．

文字単位におけるパッチ間の attention　図 3 は従来のパッチ分割と SWP を使用した場合の，ViT からなる文字符号化器における文字単位の attention の可視化である．結果は精度が最も良かった分割数のものを使用した．従来のパッチ分割の attention の可視化から，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，英数字では文字領域以外に 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存在するため，attention が分散しやすいからだと考えられる．

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．我々は実験的に 1 つあたりのパッチの特徴量を増やす SWP を提案したが，従来のパッチ分割のほうが文字の構成要素間の関係性を捉えていることが分かった．

今後は近年発展を遂げている事前学習済みモデルとの比較や，文字画像を元に文字の形状的特徴を考慮したマスク言語モデルの開発を検討している．

参考文献

[1] Yaming Sun, Lei Lin, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Radical-enhanced chinese character embedding. In Chu Kiong Loo, Keem Siah Yap, Kok Wai Wong, Andrew Teoh, and Kaizhu Huang, editors, **Neural Information Processing**, pp. 279–286, Cham, 2014. Springer International Publishing.

[2] Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao 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)**, pp. 594–598, Beijing, China, July 2015. Association for Computational Linguistics.

[3] Galina Vorobeva and Victor Vorobev. A method of the analysis of kanji structure: A new approach based on structural decomposition and coding. **NINJAL Research Papers**, No. 9, pp. 215–236, Jul 2015.

[4] Minh Nguyen, Gia H Ngo, and Nancy F Chen. Hierarchical character embeddings: Learning phonological and semantic representations in languages of logographic origin using recursive neural networks. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, 2019.

[5] Daiki Shimada, Ryunosuke Kotani, and Hitoshi Iyatomi. Document classification through image-based character embedding and wildcard training. In **2016 IEEE International Conference on Big Data (Big Data)**, pp. 3922–3927, 2016.

[6] Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. Learning character-level compositionality with visual features. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2059–2068, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[7] Shunsuke KITADA, Ryunosuke KOTANI, and Hitoshi 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.

[8] Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. Glyce: Glyph-vectors for chinese character representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 32. Curran Associates, Inc., 2019.

[9] Takumi Aoki, Shunsuke Kitada, and Hitoshi Iyatomi. Text classification through glyph-aware disentangled character embedding and semantic sub-character augmentation. In **Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop**, pp. 1–7, Suzhou, China, December 2020. Association for Computational Linguistics.

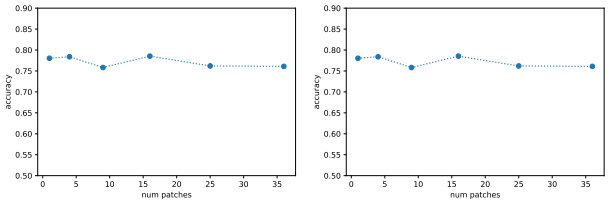
[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **International Conference on Learning Representations**, 2021.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

[13] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4190–4197, Online, July 2020. Association for Computational Linguistics.

[14] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. **Journal of Machine Learning Research**, Vol. 9, No. 86, pp. 2579–2605, 2008.



(a) 従来のパッチ分割　(b) SWP

図 4: **ViT + CLT のパッチ分割数を変化させたときの精度の変化**: パッチ分割数と文書分類の精度に大きな変化が確認できないことから，文書分類器はパッチ分割数にロバストだと考えられる．

表 2: **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)	諢	誤	誦	諍	護
ViT + SWP (25)	簣	畢	礫	鏗	謌

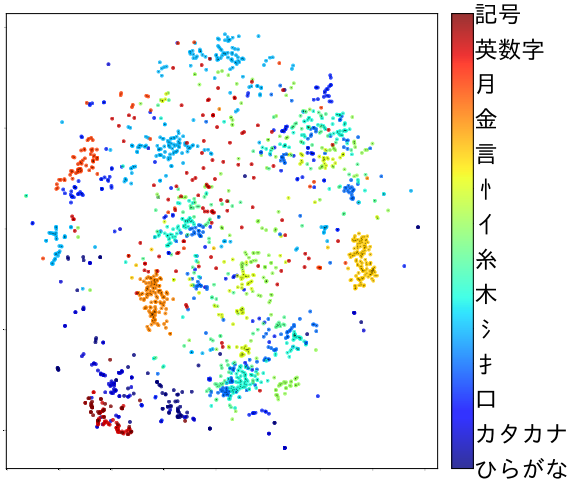
A　パッチの分割数を変化させた時の予測性能の変化

図 4 は従来のパッチ分割と提案手法 SWP によるパッチ分割において，パッチ数を変化させたときの予測性能の変化を示す．文字画像に対して分割するパッチ数の大小に関わらず，一定の予測性能を示した．この結果は従来のパッチ分割および提案手法両者ともにパッチの分割数にロバストであることが示唆されている．

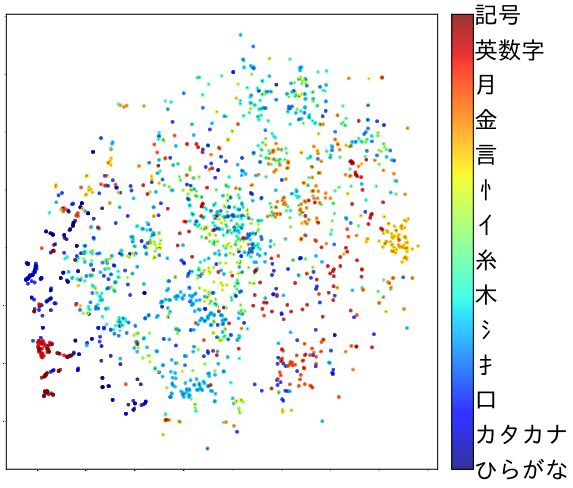
B　文字埋め込みの分析

文字埋め込みの可視化　図 5 はパッチ分割手法に従来のパッチ分割と SWP を使用した場合の ViT の文字符号化器を使用して文字画像から得られた文字埋め込みに対して t-SNE [14] で 2 次元空間へ写像して可視化した結果である．可視化の際に，その偏を持つ漢字の数が上位 10 位に入る漢字と記号や英数字等に絞って可視化した．従来のパッチ分割手法を用いた文字埋め込みの可視化から，漢字の偏を考慮した埋め込みができている．SWP を用いた文字埋め込みの可視化から，同じ構造の漢字は近傍に埋め込まれているが，部首を考慮した埋め込みはできていないことが確認された．

対象文字埋め込みの近傍の文字　表 2 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．



(a) 従来のパッチ分割 (パッチ数 = 16)



(b) SWP (パッチ数 = 25)

図 5: **主要な偏を持つ漢字と記号の埋め込みを t-SNE を用いて可視化した結果**: 従来のパッチ分割は文字の構造と漢字の偏を考慮した埋め込みが学習されており，SWP は文字の構造のみを考慮した埋め込みが学習されている．

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．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．