Free Donut: E2E 文書理解モデルにおける Attention を用いた文字領域アノテーション不要なテキスト検出手法の提案

Shuhei Yokoo[1*] Geewook Kim[2*] Sukmin Seo[3]
Atsuki Osanai[1] Yamato Okamoto[1] Youngmin Baek[1,3]
[1]LINE [2]NAVER [3]NAVER Cloud
{shuhei.yokoo,atsuki.osanai,yamato.okamoto}@linecorp.com {gw.kim,sukmin.seo,youngmin.baek}@navercorp.com

概要

本稿では，End-to-End（E2E）文書理解モデルをベースとした，文書画像から E2E にテキスト抽出と言語処理を行う新しいモデルを提案する．従来のモデルは OCR モデルや文書認識モデルといった複数のモデルを組み合わせる必要があったが，E2E 文書理解モデルはあらゆる言語処理タスクを単一モデルで扱えて，学習コストを削減することが可能である．一方で，E2E 文書理解モデルは明示的な文字検出を行わないため，テキストの位置情報の獲得ができないという問題がある．そこで，テキスト領域に Attention の注視点が分布するといった特性を利用して，位置アノテーションフリーにテキストの位置情報を獲得する方法を提案する．実験では，提案手法がくずし字認識タスクにおいて高精度な文字認識および位置情報の獲得が可能なことを示した．

1 はじめに

文書（Document）は生活のあらゆる場面で利用される．文書を自動で処理する自動文書処理（Automated Document Processing）のニーズは極めて高く，機械学習に基づく視覚的文書理解（Visual Document Understanding, 以下 VDU）モデルが提案されている．既存の VDU モデルは，テキストの検出（Detection），認識（Recognition），解析（Parsing）の 3 つのモデルを組み合わせるため，モデル学習及びアノテーションのコストが大きく，特にテキストの位置を人手で記録するコストが課題視されている．

学習コスト削減のために，単一モデルで VDU を E2E に実施するモデルが提案されている．E2E 文書理解モデルは，画像と画像から抽出したいテキストのみで学習でき，位置の教師データを必要としないのでコストが大幅に削減できる（図 1 参照）．また，複数のモデルを組み合わせる既存手法とは異なり，一つのモデルを E2E に学習するだけで VDU を実現できるため学習コストも削減できる．

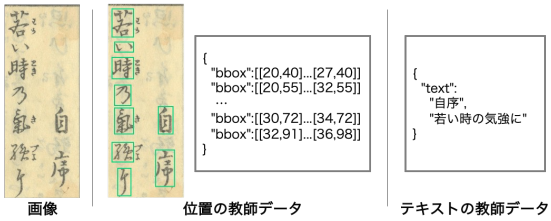

図 1 アノテーションの例. 既存手法ではテキストと位置アノテーションの両方が必須となるが，E2E モデルは位置アノテーションを必要としないためコストが大幅に削減可能.

E2E 文書理解モデルは文書分類（Classification），情報抽出（Information Extraction），質疑応答（Question Answering）などのアプリケーションで高い性能を達成したと既存研究 [1, 2] より報告されている．しかし，VDU アプリケーションには，例えば文書画像から個人情報に該当するテキストをマスキングするといった，テキストの位置が必要となる場合もある．

本研究では，テキスト位置の教師データを必要としない利点を保ちつつ，E2E 文書理解モデルによってテキスト位置を獲得する手法を提案する．モデルがテキストを出力するとき，画像上で対応するテキスト領域に注視点が分布すると仮定して，Transformer から得られる Attention Map からテキスト位置を獲得した．実験では Kim et al. [1] のモデルを用いて，古文書の画像からくずし字を検出して認識するタスク [3] に対して，最新の OCR モデル [4] との比較により有効性を検証した．

2 関連研究

光学的文字認識 (OCR). 画像からテキストを抽出する OCR は 2 つの機能で構成される．一つ目は画像上

* Equally contributed. Correspondence to Geewook Kim

の全テキストの位置情報を獲得するテキスト検出（Text Detection），二つ目は画像上のテキストを読むテキスト認識（Text Recognition）である．深層学習の発展に伴い，それぞれの機能を実現する手法が提案されている [5, 6]．また，OCR の 2 つの機能を単一モデルで E2E に実施する試みも増えている [4]．

視覚的文書理解 (VDU). VDU はテキストの抽出に加えて理解も実施する．例えば，文書画像上の情報を構造化して抽出する視覚的文書情報解析（Visual Document Parsing）や，文書画像に対する自然言語の質問文に対して回答する視覚的文書質疑応答（Visual Document Question Answering）が挙げられる．これらは主に OCR と自然言語モデルを直列に結合する手法で実現されてきた [7, 8, 9]．しかし，OCR を含む VDU モデルは，OCR で生じる計算コストの課題や，OCR まで誤差伝播しにくい学習の課題を抱えており，OCR を介さず単一モデルで VDU を実施する E2E 文書理解モデルも研究されている [1, 2]．

3 提案手法

本節ではまず E2E 文書理解モデルの一つである **Document Understanding Transformer** [1] (以下，**Donut**) の構造を説明し，Donut のような Transformer 基盤の E2E 文書理解モデルの Attention 層からテキストの位置を獲得する提案手法を説明する．

3.1 事前知識：Donut

Donut は Transformer 基盤のエンコーダ・デコーダモデルである．Kim et al.〜[1] は Swin Transformer [10] をエンコーダに，BART [11] をデコーダに用いた．エンコーダは文書画像 $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ を入力とし，画像の各パッチ領域に対応する分散表現 $\{\mathbf{z}_i|\mathbf{z}_i\in\mathbb{R}^d, 1\le i\le n\}$ を出力する．ここで d は分散表現の次元数，n はパッチ数を表すハイパーパラメータである．ここでパッチ数はまた別のハイパーパラメータである，パッチサイズ p（ここでは正方形のパッチを前提に説明する）から決まる．入力画像を $H \times W$ へ固定するとすると，パッチ数は $n = H\cdot W/p^2$ となる．デコーダはエンコーダが出力した分散表現 $\{\mathbf{z}\}$ を参照しながら求められる情報 $(\mathbf{y}_i)_{i=1}^m$ を出力する．ここで $\mathbf{y}_i\in\mathbb{R}^v$ は i 番目のトークンを表す 1-hot ベクトルであり，v は辞書のサイズ，m は出力する最大トークン数を表すハイパーパラメータである．デコーダは Cross Attention [12] と呼ばれる仕組みにより $\{\mathbf{z}\}$ を参照する．その過程でデコーダが各 time step i で参照したパッチ領域を表すスコア $\{\mathbf{A}_i|\mathbf{A}_i\in\mathbb{R}^{n\times h}, 1\le i\le m\}$（以下，Attention Map）が獲得できる．ここで h は Attention の数を決めるハイパーパラメータであり，$h > 1$ の場合は Multi Head Attention [12]（MHA）と呼ばれる．

3.2 提案アルゴリズム

提案手法は 3.1 節で説明した Cross Attention の Attention Map $\{\mathbf{A}\}$ を用いる．くずし字認識タスク [3] に転移学習した Donut モデルの Attention Map を図 2 に示す．図 2 でデコーダの最終層の h 個の Attention Map の平均値であり，値が大きい領域（以下，注視点）が各 time step で出力した文字に対応する画像上の文字周辺に集中している．デコーダの各層ごとに Map が得られるが，くずし字認識タスクに対する予備実験の結果から最終層と出力文字との対応関係が強いことが観測できた．本稿の実験と分析では最終層の Map を分析の対象とし，Text Localization へ用いる方法を考える．この傾向を用いて，各 time step の注視点をテキスト位置として獲得する．Map を box 化する方法としては，画像処理を用いることで実現した．具体的には，Map に対して連結成分ラベリングを行い，最も連結成分の値の和が大きいものを予測の box とした．さらに，我々が観測した Attention Map の 2 つの傾向を用いて Localization の精度を向上させる．

分散情報の利用． MHA を用いる Donut にはデコーダの各層に 16 個の Attention Head が存在して，16 個の Map を獲得できる．図 3 は分散が大きい Map と分散が小さい Map を示し，分散が大きい Map では注視点が出力文字周辺に集中して，分散が小さい Map では注視点が画像全体に分散する傾向を観測した．そこで，Attention Map を分散により重みつき平均した Map を最終的に用いた．

読み順の利用． 図 4 は文字出力時の Attention Map を示す．くずし字データセット [3] ではテキストの読み順が右上から左下の画像が多く，読み順に沿って文字を出力するよう転移学習した Donut では，注視点が文字の中心よりも上方に分布しやすい傾向にあった．一方で，逆方向の読み順で文字を出力するよう学習させた場合，注視点が文字の中心よりも下方に分布しやすい傾向にあった．双方向の出力が相補的な役割を果たすと期待して，これらを統合して用いる．

双方向の出力の統合と効率化について． 双方向の

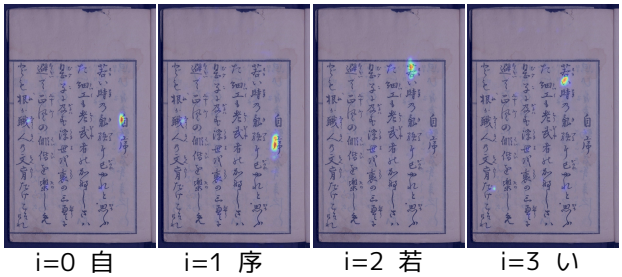

図 2　Donut モデルの Attention の可視化. くずし字認識タスクへ転移学習したモデルの Cross　Attention Map を可視化した．文字の読み順に従って Attention の注視点が移動していることがわかる．

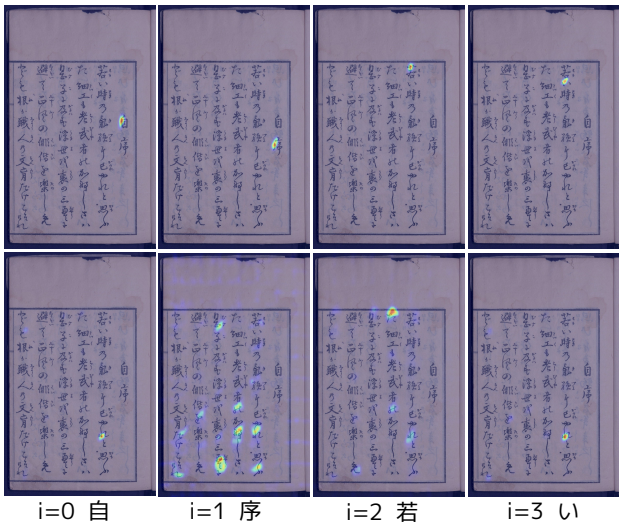

図 3　分散による Attention Map の傾向.（上）分散が最も大きい Attention Head の Map.（下）分散が最も小さい Attention Head の Map.

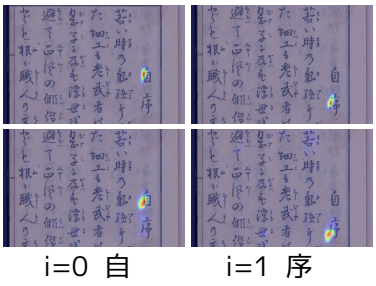

図 4　読み方向による Attention Map の傾向.（上）順方向の Map.（下）逆方向の Map.

出力を統合する方法には, 双方向の Attention Map の重みつき平均を算出する方法（以下，Early Fusion）と, 双方向の Attention Map からテキスト位置を示す Bounding Box を獲得して，それらを統合する方法（以下，Late Fusion）が考えられる．Box を統合する手法としては，WBF [13] を用いた．本稿では両手法を実装して 4 節で定量評価した．

次に，提案手法の学習と推論の効率性について考察する．Donut は次のステップの文字を予測させる

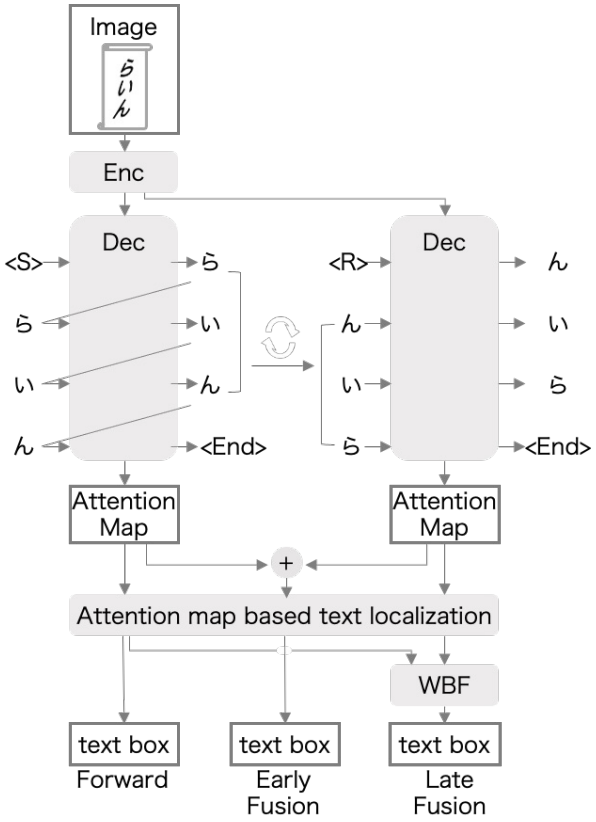

図 5　提案手法の Overview. テキスト出力時の Atteniton Map から位置を獲得する．<S>は順方向の文字出力, <R>は逆方向の文字出力を動作させる命令トークン [1] である．

Teacher Forcing 方式で学習する [1]．逆方向の文字出力は，順方向のテキストを反転させたテキストを教師データに用いて学習すれば容易に実現できる．また，そのとき双方向モデルの重みを全て共通化することで，パラメータ数を増加させることなく学習可能である．さらに，双方向どちらのデコーダも，エンコーダの同一の出力を入力に用いるため，エンコーダの計算は一回で済むことも利点として挙げられる．

推論時は 1 ステップに 1 文字づつ出力するため，双方向で計 2 回の推論を実施すると計算コストが大きくなってしまう．そこで次のアイデアを適用した．まず，片方向（ここでは順方向と仮定する）だけ推論して Attention Map と文字列を得る．次に，得た文字列を反転させた文字列を，逆方向の読み順でモデルが出力した文字列だと見なして，全て同時にデコーダに入力して逆方向の Attention Map を得る．この方法だと，逆方向の推論時にステップを繰り返すことなく，1 ステップのみで完了するため計算コストが大幅に削減できる．図 5 に提案手法の全体フローを示す．

Method	mAP@.5	mAP@.3	mAP@.1	CLEval@.5	CLEval@.3	CLEval@.1	F1	nED
MS Azure (API)	0.056	0.201	0.227	0.134	0.134	0.134	0.374	0.746
EasyOCR (OSS)	0.004	0.006	0.007	0.000	0.000	0.000	0.017	0.978
DEER (Trained)	**0.883**	0.883	0.884	**0.950**	**0.952**	**0.953**	0.902	0.053
Free Donut (Proposed, Late-fusion)	0.795	**0.901**	**0.914**	0.674	0.872	0.940	0.933	0.035
Free Donut (Proposed, Early-fusion)	0.769	0.896	0.913	0.638	0.843	0.919	**0.966**	**0.029**
Free Donut (Forward Only, Not Weighted)	0.472	0.868	0.909	0.583	0.725	0.772	0.961	0.030
Free Donut (Forward Only)	0.667	0.886	0.910	0.591	0.810	0.888	0.958	0.031
Free Donut (Reversed Only)	0.693	0.889	0.905	0.679	0.867	0.933	0.901	0.037

表 1　くずし字認識タスクの性能.

4　実験

タスクとデータ. 実験では古文書くずし字データセット [3] 用いて文字検出及び認識タスクにより提案手法を評価した．データセットに含まれる 44 冊の古文書から一冊を評価データとして用いた．本タスクではくずし字が書かれている文書画像から各くずし字の位置とテキストを獲得しており，くずし字を現代の語彙へと認識する点で従来の OCR タスクとは若干異なる点に注意されたい．

評価方法. 評価指標には検出した文字の位置を評価する mAP，OCR の評価指標の 1 つである CLEVAL [14]，くずし字認識 Kaggle チャレンジ [1]で用いられた F1 スコア，そして認識した全ての文字を検出位置を元にソートして，Ground Truth の文字列との一致度合いを評価する Normalized Edit Distance（nED）を用いて評価した．

比較手法. まず，本研究では終端間文書理解モデルの一つである Donut [1] に提案手法を結合したものを提案手法とし，Box Annotation Free Donut（Free Donut）と呼ぶ．比較対象には，利用可能な API とオープンソースモデルである，MS OCR API[2]，EasyOCR[3]を使用した．くずし字認識に特化した最先端手法とも比較するために，近年開催された ECCV-22 OCR Challenge [4]を参照し，DEER [4] をくずし字データセットで学習した特化モデルも準備した．

結果. 表 1 に実験結果を示す．まず上段は市販の API やオープンソースモデルの評価結果であり，他の手法より大幅に低いスコアとなった．これはくずし字認識に特化していないためだと推察される．次に，中段はくずし字認識タスクで学習したモデルの評価結果である．テキスト位置の教師データを用いた DEER [4] と比べて，テキスト位置の教示データを用いない提案手法でも一部の評価指標では同等以上の性能を達成できた．ただし，出力した Bounding Box と Ground Truth Box の一致度合いを表す Area Precision [14] では，閾値次第で DEER とのスコアの差分が大きくなった．これはテキスト位置の教師データを用いて学習した DEER に比べて，テキスト位置の教師データを用いない提案手法では，テキスト領域を余白なく漏れなく特定する観点では劣ることを示唆し，改善の余地がると分かった．出力した Bounding Box の中心点が，GT の Bounding Box 内に含まれるかどうかを評価する F1 や，認識されたテキストの正しさを評価する nED では提案手法が優位となった．

次に，Ablation Study により提案手法を構成する 2 つの要素の有効性を評価した．まず，分散を用いて Attention Head ごとの Map の重みつき平均を取ることによる性能向上は下段の 1 行目と 2 行目の比較で確認できる．次に，双方向の読み順の出力を利用することの有効性は，中段の提案手法のスコアと，下段の片方向のみを利用した場合のスコアとの性能差から確認できる．

5　まとめと今後の課題

本稿では E2E 文書理解モデルを用いて，推論時の Attention Map の後処理によってテキスト位置を獲得するアルゴリズムを提案した．実験では古文書のくずし字認識タスクで高いスコアを示した．今後の課題は，Attention Map を後処理する Ad-hoc なアルゴリズムの正式な根拠づけ，性能向上，そして文字認識に限らない多様な応用先での有用性評価である．

1　https://www.kaggle.com/c/kuzushiji-recognition
2　https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/overview-ocr
3　https://github.com/JaidedAI/EasyOCR
4　https://rrc.cvc.uab.es/?ch=19&com=evaluation&task=1

謝辞

The authors deeply thank members of NAVER Cloud CLOVA Vision Semantic Perception Team and LINE AI Company Computer Vision Lab Team.

参考文献

[1] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In **European Conference on Computer Vision**, pages 498–517. Springer, 2022.

[2] Brian Davis, Bryan Morse, Bryan Price, Chris Tensmeyer, Curtis Wigington, and Vlad Morariu. End-to-end document recognition and understanding with dessurt. **ECCV Workshop on Text in Everything**, 2022.

[3] 国文学研究資料館. 日本古典籍くずし字データセット.

[4] Seonghyeon Kim, Seung Shin, Yoonsik Kim, Han-Cheol Cho, Taeho Kil, Jaeheung Surh, Seunghyun Park, Bado Lee, and Youngmin Baek. Deer: Detection-agnostic end-to-end recognizer for scene text spotting. **arXiv preprint arXiv:2203.05122**, 2022.

[5] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In **Proceedings of the IEEE/CVF international conference on computer vision**, pages 4715–4723, 2019.

[6] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pages 9365–9374, 2019.

[7] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, pages 1192–1200, 2020.

[8] Teakgyu Hong, DongHyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. **Proceedings of the AAAI Conference on Artificial Intelligence**, 36(10):10767–10775, Jun. 2022.

[9] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. **arXiv preprint arXiv:2204.08387**, 2022.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pages 10012–10022, October 2021.

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[13] Roman A. Solovyev and Weimin Wang. Weighted boxes fusion: ensembling boxes for object detection models. **arXiv preprint arXiv:1910.13302**, 2019.

[14] Baek Youngmin, Nam Daehyun, Park Sungrae, Lee Junyeop, Shin Seung, Baek Jeonghun, Young Lee Chae, and Lee Hwalsuk. Cleval: Character-level evaluation for text detection and recognition tasks. **arXiv preprint arXiv:2006.06244**, 2020.