N-012

日本近代公文書 OCR における画像特徴抽出器の比較

Comparision of Image Feature Extractors in Modern Official Document OCR

宮川 裕貴† 山田 雅之† 中 貴俊† 兼松 篤子†
Yuki Miyagawa Masashi Yamada Takatoshi Naka Atsuko Kanematsu
宮崎 慎也† 長谷川 純一‡
Shinya Miyazaki Junichi Hasegawa

1. はじめに

日本の各行政機関が保管する公文書は政治決定の背景や当時の日本国内外の情勢を知ることができるというように史料的価値があるが，戦前期の文書の多くは近世古文書の流れを汲む手書き文字による文書であるため，解読には専門的な知識が必要である．したがって一般行政職員が解読することは容易ではなく，また，解読の知識を持つ研究者も少ないため公文書史料が活用できていないのが現状である．我々の研究グループは台湾総督府文書を題材として，OCR システムとそのためのデータセット開発を進めており，行単位のテキスト画像を対象とした文字領域検出とその文字認識を行うエンコーダ・デコーダベースのモデルを考案した．このモデルはエンコーダで画像の特徴を抽出，デコーダで文脈情報と合わせた特徴量に変換し，これを用いて文字領域の予測と文字クラス分類を行う．本稿ではエンコーダの画像特徴抽出器を変更した複数のモデルで認識精度を比較し，文書画像の特徴抽出に適したモデルを考察する．

2. 関連研究

山田らは歴史的文書の行単位画像に対する文字矩形と文字列予測手法として，画像情報，文字列情報，文字矩形情報を用いた手法を提案している[1]．また，Le らは明治初期から中期に出版された雑誌を題材として，Attention と LSTM を組み合わせたエンコーダー・デコーダベースのテキスト認識モデルを開発した[2]．

3. データセット

データセットは近代公文書データセットを使用する．これは我々の研究グループが，台湾総督府文書画像（図 1）に対して文字領域を表す矩形情報とその文字ラベルをアノテーションデータとして付与したものである．アノテーションはページ単位で与えられており，本実験ではページ単位の画像を元に，図 2 に示すような 1 行から 3 行のテキスト画像として切り出した実験用データセットを作成した．実験用データセットは 34,467 枚の画像とそのアノテーションデータで構成されており，これを訓練用：検証用：テスト用 = 8:1:1 で分割した．

図 1：台湾総督府文書画像

図 2：実験用データセットの例

4. 学習条件・評価指標

本実験で使用するモデルは大きく分けて 3 つの構造を組み合わせて構成する．1 つ目は画像の特徴抽出と画像サイズの縮小を行うバックボーンネットワーク，2 つ目は Attention 機構を用いて画像の特徴抽出を行う Transformer Encoder，3 つ目は Transformer Decoder と

† 中京大学 Chukyo University
‡ 中京大学人工知能高等研究所 Institute for Advanced Studies in Artificial Intelligence

線形層からなる構造で，これは文脈情報と抽出された画像特徴を用いて文字領域と文字クラスを予測する．実験ではバックボーンネットワークの比較対象として CNN ベースの ResNet[3] と Transformer ベースの Swin Transformer[4] を取り上げる他，Transformer Encoder の有無による精度の違いを検証する．評価指標は適合率，再現率を用いる．

5. 実験結果

表 1 は各モデルのパラメータ数と，テストデータに対する適合率と再現率を示している．ResNet の後ろに続く数字は層数を表しており，Swin は Swin Transformer の略称，TE は Transformer Encoder の略称とする．また，Swin の後ろに続くアルファベットは隠れ層の次元数や層数の違いを表しており，規模の大きいものから B，S，T となっている．表 1 に示した結果より，パラメータ数が最も多い Swin-B + Transformer Encoder が適合率，再現率において最大の結果を示した．しかし，ResNet101+TE と Swin-T+TE を見るとパラメータ数に 15M の差があるが，精度に大きな差が見られないことから，単純にパラメータ数を増やせば精度が向上するとはいえない．また，バックボーンネットワークが同一の場合，Transfomer Encoder を含むモデルが精度が高いことや，ResNet50+TE と ResNet101 の比較，Swin-T+TE と Swin-S との比較より，画像特徴抽出器の構造としてバックボーンネットワークと Transformer Encoder の両方を採用することが有効といえる．図 3 に Swin-B + Transformer Encoder の予測例を示す．青文字，青枠はそれぞれ予測した文字クラスと文字矩形を表す．多くの文字を検出できている一方，1 つの文字を複数の文字として認識しているなどの過検出が見られる他，1 文字間違えると次の文字も間違えやすい傾向にあるため，これらへの対応が必要である．

6. まとめと今後の展望

本稿では文書画像認のためのモデルにおける画像特徴抽出器を変更した場合の認識精度の違いについて述べた．実験を通じて，バックボーンネットワークに加えて Transformer Encoder の利用が有効であると判明した．今後は認識精度の向上を目指し，事前学習手法の検討や文字領域の情報を利用可能なモデルの考案を行う．

表 1 ：パラメータ数・適合率・再現率

モデル名	パラメータ数	適合率	再現率
ResNet50	54M	82.59	81.12
ResNet50 + TE	73M	89.47	89.06
ResNet101	73M	83.95	82.70
ResNet101 + TE	92M	90.06	89.80
ResNet152	89M	86.41	85.38
ResNet152 + TE	108M	91.11	90.94
Swin-T	58M	82.87	81.95
Swin-T + TE	77M	90.91	90.74
Swin-S	79M	88.61	88.03
Swin-S + TE	98M	91.30	91.04
Swin-B	118M	89.58	89.17
Swin-B + TE	**137M**	**91.88**	**91.60**

図 3：予測結果の例

謝辞

本研究は JSPS 科研費 JP20J01304, JP23K20105 の助成を受けた．

参考文献

[1]山田雅之 他,”歴史的文書データセットの文字矩形情報を用いた行単位画像からの文字列予測とセグメンテーション”，情報処理学会論文誌, Vol.65, No.3, pp.754-766, 2024

[2] Anh Duc Le, Daichi Mochihasi et al. “Recognition of Japanese historical text lines by an attention-based encoder-decode and text line generation”, HIP’19, pp.38-41, 2019

[3] Kaiming He et al. , “Deep Residual Learning for Image Recognition”, CVPR, pp.770-778, 2016

[4] Ze Liu et al., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”, ICCV, pp.10012-10022, 2021