4U-08

Transformerデコーダを用いた画像内のテキスト領域検出の検討

矢島 英明 †
山梨大学大学院医工農学総合教育部 †
レオ チーシャン ‡
山梨大学大学院医工農学総合教育部 ‡
北川 智樹 §
山梨大学大学院医工農学総合教育部 §
西崎 博光 ¶
山梨大学大学院医工農学総合教育部 ¶

1 はじめに

近年，深層学習技術の進歩は，画像認識，自然言語処理，その他の多くの分野でも画期的な成果を生み出している．特に，テキスト認識は，文書の Optical Charactor Recognition (OCR)，道路標識の認識，情報抽出など多岐にわたる応用において重要な役割を果たしている．従来の深層学習アプローチでは，深層学習モデルを用いて画像内の特徴を抽出し，後続の画像処理技術によってテキスト領域を特定するバウンディングボックスを決定することが一般的であった．

しかしながら，この方法にはいくつかの制約が存在する．例えば，画像処理に基づくバウンディングボックスの生成は，画像のノイズや複雑な背景に弱いという問題がある．また，モデルの出力と画像処理技術を用いた後処理の適用次第で，テキスト検出精度の低下を引き起こすこともある．

これらの課題に対処するため，本稿では，Transformer デコーダ [1] を用いた新しいアプローチを提案する．このアプローチでは，Vision Transformer[2] をエンコーダとして特徴を抽出し，これらの特徴を Transformer デコーダへ入力する．デコーダは，これらの特徴から直接バウンディングボックスをテキスト形式で推定する．この方法により，従来の画像処理技術を用いた後処理ステップを省略し，エンドツーエンドでのテキスト領域の検出が可能となる．

実験では，合成して生成したテキスト画像を用いてモデルの訓練と評価をした結果，高い精度でテキスト検出ができることを確認した．本手法は，画像内のテキスト領域をより正確かつ効率的に検出する可能性を秘めている．

2 モデル

本研究において使用するモデルは，Vision Transformer と Transformer デコーダを組み合わせたアーキテクチャに基づいている（図 1）．ここでは，特に Transformer-based Optical Character Recognition (TrOCR) [3] モデルを採用している．TrOCR は，Transformer アーキテクチャを活用して OCR タスクの精度を向上させるために設計されており，画像からのテキスト抽出と認識において顕著な結果を示している．

本研究におけるモデルの目的は，画像内のテキスト領域を表すバウンディングボックス座標を直接テキストとして推定することである．以下，モデルの各コンポーネントとその動作について説明する．

2.1 Vision Transformer(ViT)

Vision Transformer は，画像をパッチに分割し，それぞれのパッチを線形に埋め込むことで画像を処理する．これらの埋め込みは，位置エンベッディングと共に Transformer デコーダに入力される．

ViT は，画像の局所的および大域的な特徴を捉える能力を持ち，テキスト検出のための強力な特徴表現を提供する．

2.2 Transformer デコーダ

Transformer デコーダは，ViT からの特徴表現を受け取り，テキスト領域を表すバウンディングボックスの座標情報を直接テキストとして生成する．

デコーダは，Self-Attention と Cross-Attention を使用して，画像内の関連する特徴を識別し，それに基づいてテキスト領域を推定する．このプロセスは従来の画像処理ベースのアプローチと比較して, 直接的かつ効率的なテキスト領域の推定を実現する．

2.3 モデルのトレーニング

本研究でのモデルのトレーニングは，シーケンスツーシーケンス（Seq2Seq）タスクとして行われる．このアプローチでは，アノテーション付きの合成テキスト検出データセットを使用し，各画像に対応するバウンディングボックスのテキスト情報をシーケンスとしてモデルに学習させる．モデルは，画像から得られた特徴を基に，対応するテキスト領域のバウンディングボックスをシーケンスとして生成する能力を学習する．

トレーニングのプロセスでは，画像とそれに対応するバウンディングボックスのテキスト情報が入力として使用される．損失関数は Cross Entropy を使用している．この損失関数は予測されたバウンディングボックスのテキストシーケンスと実際のバウンディングボックスのテキストシーケンスとの間の一致度を測定し，差異を最小化する．また，トレーニングでは，適合率，再現率，F1 スコアを使用してモデルの予測精度と実際のテキスト領域との一致度を定量的に評価する．

3 実験

3.1 データセット

実験では，異なる種類の文書画像を模した合成テキスト検出データセットを作成し使用した．このデータセットは，様々なフォント，サイズ，および手書き文字とフォント文字かのスタイル区別を含む 12,500 枚の画像で構成されている．各画像には，対応するバウンディングボックスのアノテー

A Preliminary Study of Text Area Detection in an Image Using a Transformer Decoder
† Hideaki Yajima, University of Yamanashi
‡ Chee Siang Leow, University of Yamanashi
§ Tomoki Kitagawa, University of Yamanashi
¶ Hiromitsu Nishizaki, University of Yamanashi

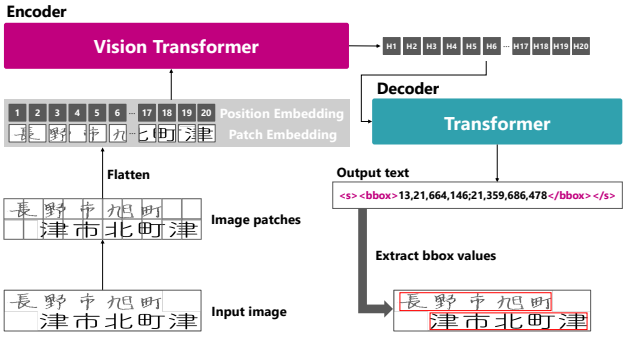

図 1　モデル概略

ションが含まれている．737 種類のフォントを用いており，手書き文字画像には ETL データベース [4] の画像を使用している．データセットの作成プロセスでは，空白の画像に 1 行から 5 行までの範囲で手描き文字画像またはフォント画像をテキストに合わせて貼り付ける方法を採用した．これは，実際の文書画像中に存在するような複数行テキストを模倣し設計したものである．

データセットを使用する際には，画像内のテキストとそれに対応するバウンディングボックスを識別し，モデルが解析可能な形式に変換するための処理を施した．この処理により，バウンディングボックスの情報は特定の構文で表現される．これにより，モデルは画像内のテキストの位置をより正確に認識し，解析することができる．このデータセットの設計は，モデルが現実世界の多様なテキスト検出シナリオに適応できるように，画像の多様性と複雑さを考慮している．

3.2　実験条件

本研究では，合成テキスト検出データセットの中から 10,000 枚を訓練用，2,000 枚を検証用，500 枚を評価用に使用した．モデルには，事前学習された ViT と Transformer デコーダを組み合わせた microsoft/trocr-base-handwritten[*1] を採用した．この事前学習モデルは，画像から手書きテキストを認識するために設計されているモデルである．

実験において，各画像はモデルへの入力前に 384×384 ピクセルにリサイズされる．ミニバッチサイズは 8，最適化関数には AdamW を使用し，学習率は 0.0005 から開始してスケジューラにより徐々に減少させた．また，損失関数には Cross Entropy を採用した．

モデルが特殊トークン `<bbox>`, `</bbox>` を認識できるようにトークナイザーのカスタマイズを行った．一枚の画像に複数のテキストが存在する場合，

$$\texttt{<bbox>}x_{11}, y_{11}, x_{12}, y_{12}; x_{21}, y_{21}, x_{22}, y_{22}\texttt{</bbox>}$$

のようにセミコロン「;」をデリミタとして使用し，各テキストに対するバウンディングボックスの座標を表現した．そして，それらのテキスト単位のバウンディングボックスの座標を推定させるというアプローチの有効性を評価した．

評価指標には，Precision（適合率），Recall（再現率），および F1 スコアを含む複数の評価指標を使用した．これらの指標は，予測されたバウンディングボックスと実際のバウンディングボックスとの間の類似性と精度を測定するために重要である．

[*1] https://huggingface.co/microsoft/trocr-base-handwritten

表 1　評価データに対する実験結果 (IoU= 0.75)

Precision	Recall	F1
0.8763	0.8673	0.8718

3.3　結果

表 1 に評価データにおける各評価指標の結果を示す．実験の結果，本研究におけるモデルはテキスト領域の検出において高い精度を達成した．特に，複雑な構成や異なるテキストスタイルが存在する画像においても，バウンディングボックスの正確な推定が可能であることが示された．

これらの結果は，ViT と Transformer デコーダの組み合わせがテキスト検出タスクにおいて有効であることを示唆している．

4　おわりに

本研究では，ViT と Transformer デコーダを組み合わせたテキスト検出モデルの新たなアプローチを提案し，その有効性を実験的に検証した．このアプローチにより，画像内のテキスト領域を効率的かつ正確に検出することが可能であることが示された．特に，複雑な文字サイズや異なるテキストスタイルを含む画像においても，高い精度でテキスト領域を識別できることが実証された．

本研究におけるアプローチは，従来の画像処理ベースのアプローチと比較してバウンディングボックスの生成において，画像のノイズや複雑な背景の影響を受けにくい可能性があり，また，end-to-end での処理により，処理の効率化と精度の向上が考えられる．

これらの特性は，特に OCR や情報抽出などの応用において重要な意味を持つ．さらに，本研究で使用された合成テキスト検出データセットは，現実世界の多様なシナリオを基にしており，モデルの一般化能力の評価にも貢献している．しかし，実際の応用を考慮すると，より多様なデータセットや実世界のデータを用いたさらなる検証が必要である．

将来的には，モデルのさらなる改良や最適化，さらには異なる種類のデータセットへの適用を通じて，このアプローチの潜在能力をさらに探求していくことが重要である．

今後の展望として，提案モデルの精度向上に寄与する既存手法との統合や共生により，さらなるテキスト検出精度の向上を目指す．

参考文献

[1] A. Vaswani, et. al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems, vol. 30, 2017.
[2] A. Dosovitskiy, et. al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Proc. of ICLR, pp. 1–21, 2021.
[3] M.Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, “Trocr:Transformer-based optical character recognition with pre-trained models.” Proc. of AAAI, pp.13094–13102, 2021.
[4] 独立行政法人産業技術総合研究所, “ETL 文字データベース” http://etlcdb.db.aist.go.jp, (2024.1.10 参照)