手書き文字認識モデルエンコーダへの言語特徴付与による有効性

工藤 滉青（922022）　　指導教員：瀬尾 昌孝

1. はじめに

本研究では,手書き文字認識（Handwritten Text Recognition: HTR）を対象とする.HTR は,手書き文字画像から対応する文字列を自動的に認識する技術である.既存の HTR モデルでは視覚特徴取得のみに基づいたエンコーダが中心であり,字形の曖昧さや画像の欠損・損傷に対して誤認識が生じやすい.そこで,Connectionist Temporal Classification(CTC)に基づく認識結果に対して外部の言語モデルを用いたデコードを行うことによって,文全体の構文的一貫性や語彙的整合性といった言語的側面を補完する手法が広く用いられている.一方で,言語知識は主に推論時の補正として利用され,学習段階でエンコーダの特徴抽出能力に十分反映されていないという課題がある.

本研究では,LLM による中間損失を導入し,文全体の構文的一貫性や語彙的整合性といった言語的側面を,視覚特徴と同時にエンコーダで学習可能な枠組みを提案する。

2. 関連研究

2.1 TrOCR

TrOCR[1]は Transformer を用いた HTR モデルであり,エンコーダに ViT 系モデル,デコーダに言語モデルを採用した Encoder–Decoder 構造を持つ.TrOCR では,従来 CTC と外部言語モデルの組み合わせによって補われていた言語的知識を,デコーダ内部に組み込むことで外部言語モデルを不要としている.

本研究では,LLM に基づく中間損失の効果を明確に検証するため,TrOCR のデコーダを用いず,CTC デコーダに置き換えた TrOCR-CTC を採用する.言語的制約の弱い CTC を用いることで,中間損失による言語知識注入の寄与をより直接的に評価可能とする.

2.2 Language-Aware Intermediate Loss

Language-Aware Intermediate Loss(LAIL)は,LLM が持つ言語的知識を活用して,CTC ベースの音声認識モデルを強化する補助損失手法である.Conformer エンコーダの複数の中間層出力をコネクタ層によって LLM の埋め込み空間へ写像し,各層において CLM(causal language modeling)損失を計算する.学習時には,CTC 損失と LAIL を組み合わせて最適化することで,エンコーダの中間表現に言語的知識を段階的に注入する.この手法により,非自己回帰型デコーディングの高速性を維持しつつ,文全体の構文的・語彙的整合性を考慮した特徴表現の獲得が可能となる.本研究では,この LAIL の枠組みを手書き文字認識に拡張する.

3. 提案手法

本研究では,LLM に基づく中間損失の効果を明確に検証するため,TrOCR のエンコーダを基盤とし,デコーダを CTC に置き換えた TrOCR-CTC を採用する.CTC は言語的制約が比較的弱く,中間損失による言語知識注入の寄与を他要因から分離して評価しやすいという利点がある.提案モデルでは,TrOCR エンコーダの複数の中間層出力にコネクタ層を接続し,各中間特徴を LLM の埋め込み空間へ写像する.写像後の特徴ベクトルを凍結した LLM に入力し,正解文字列に対する CLM 損失を中間損失として計算する.これにより,エンコーダは CTC による文字識別に加え,文全体の構文的・語彙的整合性を反映した表現を獲得するように学習が促される.学習時の損失関数は,CTC 損失と LAIL を加算したものとし,LAIL は複数の中間層に付与する CLM 損失の重み付き和として定義する.提案手法の全体構造を図 1 に示す.

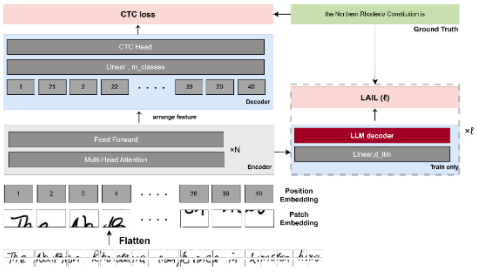

図 1 提案手法モデル

4. 実験

提案手法の有効性を検証するため, LLM を用いないベースラインモデル（TrOCR-CTC）と，LLM に基づく中間損失を導入した提案モデルの性能比較を行った．評価指標には CER を用いた．ベースラインと提案モデルの比較結果を表 1 に示す.

表 1 ベースラインモデルと提案手法モデルの結果

ベースライン(CER:%)	提案モデル(CER:%)
13.64	**10.77**

5. 結論

本研究では，HTR において LLM に基づく中間損失を導入し，学習段階から言語的知識をエンコーダ表現へ反映させる手法を提案した．実験の結果，提案モデルはベースライン（TrOCR-CTC）に対して CER を 13.64%→10.77%に改善し，本手法の有効性を確認した．今後はコネクタ層の設計の最適化と，異なるデータセットへの汎化性能評価を行う.

参考文献

[1] Li et al.,“TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models,” AAAI, 2023.