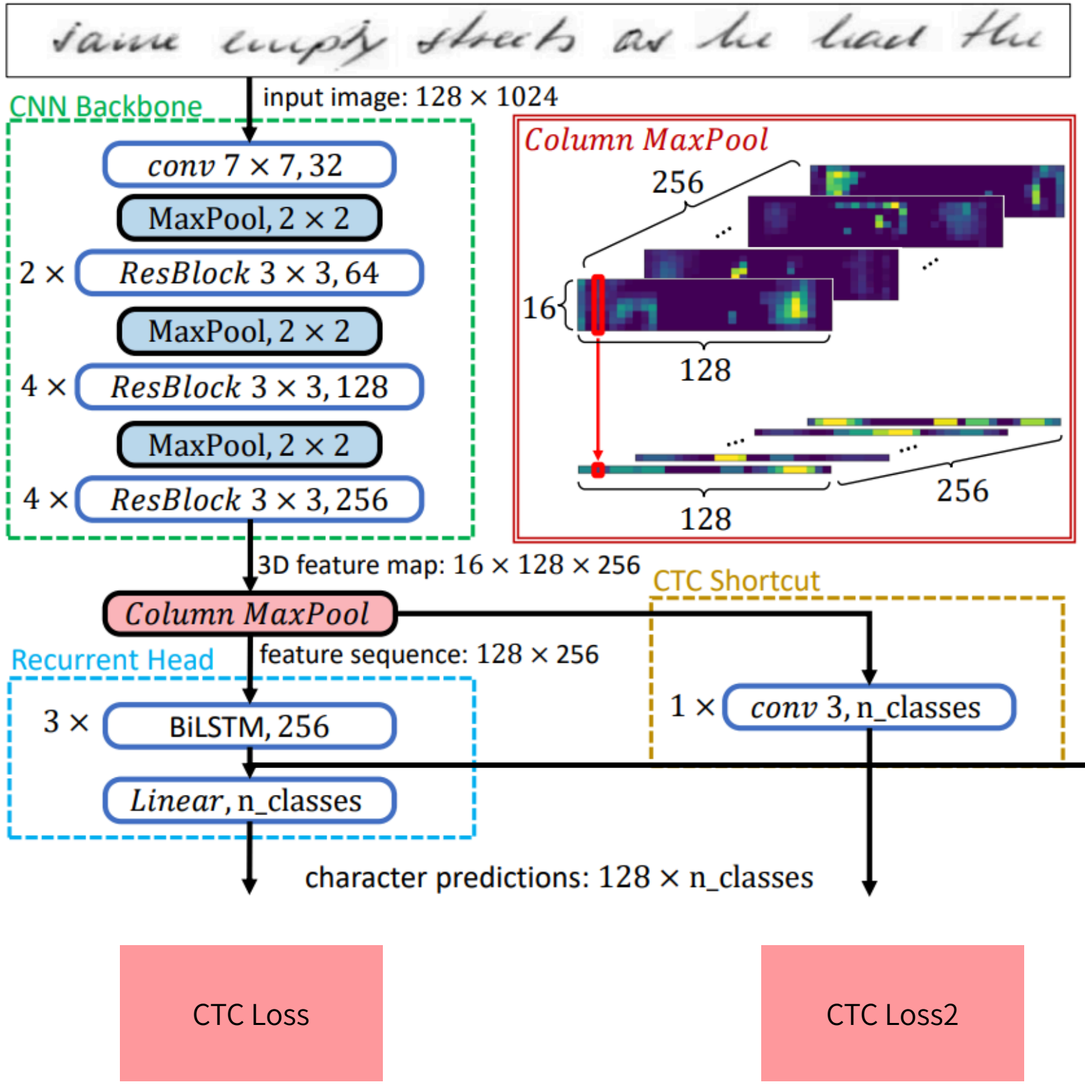
LLM条件付き確率最大化に基づく手書き文字認識の精度向上

2025/10/15

提案モデル



connector layer → **LLM (Llama 3)** → LAIL

前回のゼミでの質疑応答(一部抜粋)

文脈情報をと入れたいならtransformerでよいのでは？
→少量データで精度向上目指したい
→transformerベースのモデルではattentionの構造上計算効率が悪く大量の合成データが必要

少量データでの精度向上は何がいいの？？
→昨今、生成モデルなどの発展により今回のようなデータの合成、生成難易度はそんなに高くないよね
→answer　：タイムアップ！！

少量データで学習できるメリット

そもそも

- 画像データ増やすこと自体のそんなに難易度って高くないよね
  - 生成モデル等による生成方法の確立されてきつつある
- 少量データの規模間
  - 数千行データ

少量データで学習できるメリット

メリット

1. 学習効率がいい
2. **希少ドメインに適用可能（収集が高コスト／倫理的に絞られる領域）**

有用なタスク例

- 個人化（ユーザー固有の癖・環境に数百〜数千で素早く追従）

結果

Epoch	Val CER	Val WER	Test CER	Test WER
800	**0.032**	**0.116**	0.046	0.157
700	**0.032**	**0.116**	0.046	**0.156**
600	**0.032**	**0.116**	0.046	0.158
500	0.033	0.116	0.047	0.158
400	0.038	0.134	0.052	0.175
300	0.038	0.132	0.054	0.178
200	0.04	0.14	0.056	0.191
100	0.04	0.143	0.057	0.194
50	0.044	0.154	0.064	0.21

CRNN-only

Epoch	Val CER	Val WER	Test CER	Test WER
800	**0.033**	0.118	**0.045**	0.155
750	**0.033**	0.117	**0.045**	0.155
700	**0.033**	0.118	**0.045**	**0.153**
650	**0.033**	0.118	**0.045**	0.154
600	0.034	0.12	0.046	0.157
550	**0.033**	**0.116**	**0.045**	**0.155**
500	0.034	0.12	**0.046**	0.156
450	**0.033**	0.119	**0.046**	0.157
400	0.038	0.134	0.053	0.177
350	0.04	0.138	0.051	0.175
300	0.038	0.133	0.054	0.183
250	0.04	0.136	0.056	0.187
200	0.039	0.138	0.054	0.183

Proposed

$$\mathrm{CER} = \frac{\text{置換} + \text{挿入} + \text{削除}}{\text{正解の総文字数}}$$

原因と対策

- LLM は prefix がなくても「過去の tok_emb（＝正解文字列の埋め込み）」だけで次トークンを高精度に予測しちゃうやん！！！！！！！
  例：y4の予測時　[z1,z2,...zl,y1,y2,y3,mask1,mask2,...maskn]
  z(中間表現)を参照せずに次トークン予測してるんじゃね💦

対策

- PrefixのみでLLMの入力にする？？
  - さすがにコネクタ層も未学習だし学習が不安定になりそうだ
- 正解文字列のマスク？？
  - 段階的にマスクしてみる
  - 最後にはprefix-only で