

2026.02.06 令和7年度卒業論文審査会

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

大阪工業大学 ロボティクス&デザイン工学部 システムデザイン工学科

学生番号：922022

氏名:工藤滉青

指導教員：瀬尾昌孝

アウトライン

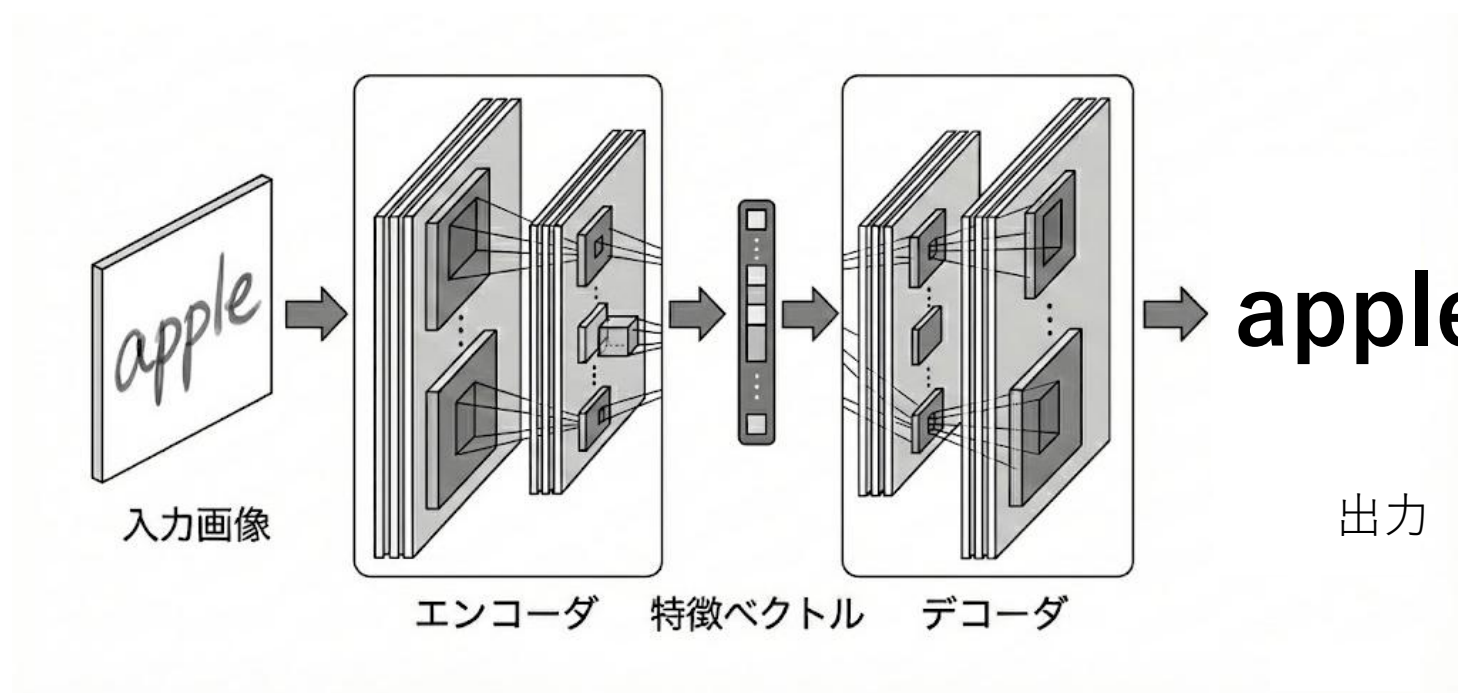
1. 研究テーマ
2. 関連研究
 - 2.1. Transformer-based OCR(TrOCR)
 - 2.2. LLM損失によるエンコーダ学習
3. エンコーダへの言語特徴付与
4. 実験
5. まとめ

研究テーマ

手書き文字画像認識の問題点

課題：

画像を取り扱うエンコーダでは**視覚特徴**に依存



研究テーマ

手書き文字画像認識の問題点

課題：

画像を取り扱うエンコーダでは**視覚特徴**に依存

→字形曖昧時での誤認識が発生



cam**se** wlat may



he w**o**s l**o**wyer

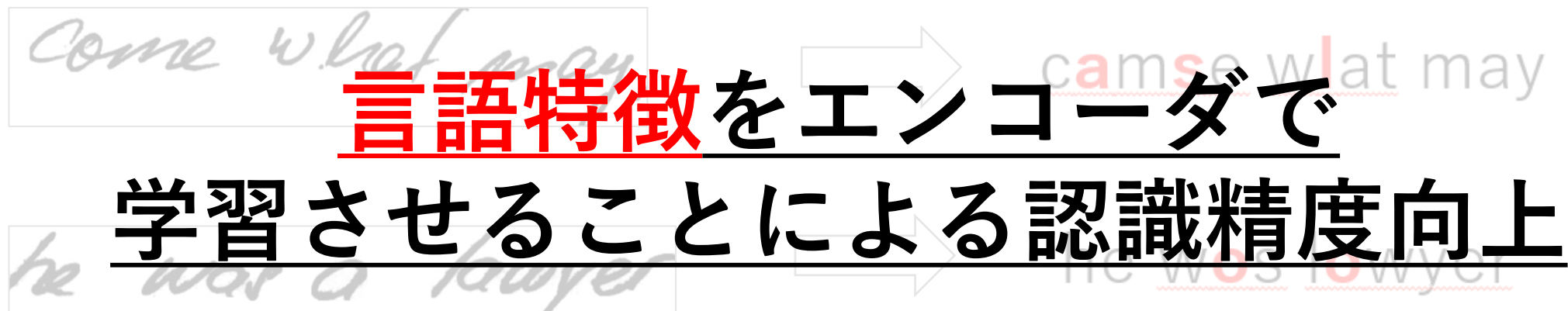
研究テーマ

手書き文字画像認識の問題点

課題：

画像を取り扱うエンコーダでは**視覚特徴**に依存

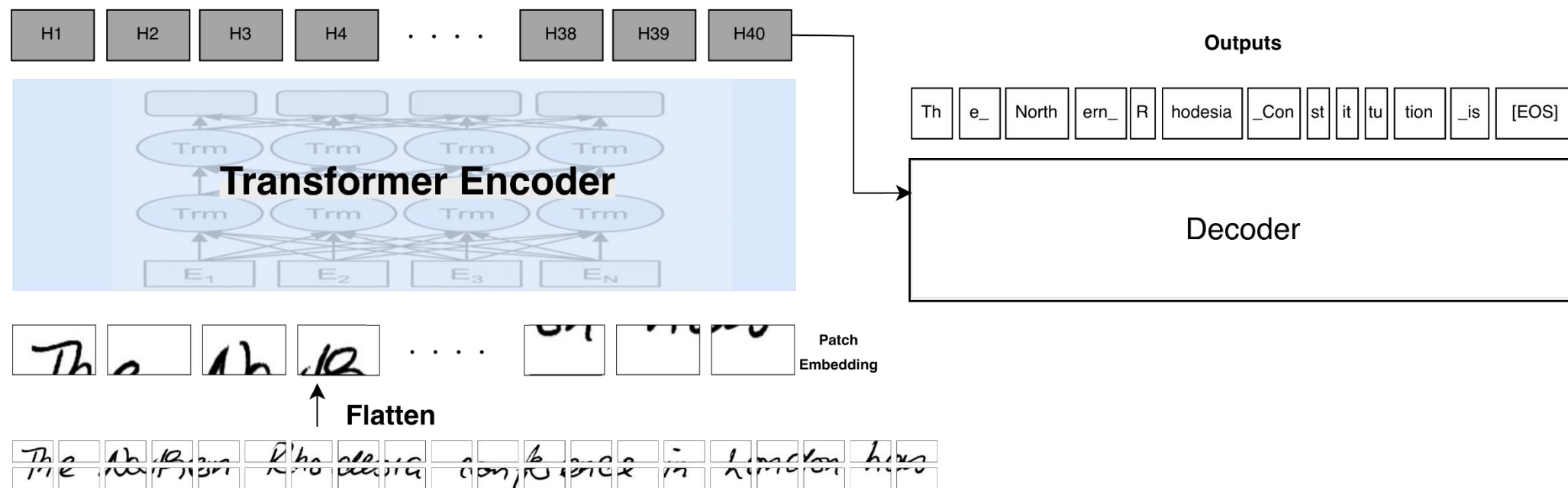
→字形曖昧時での誤認識が発生



言語特徴をエンコーダで
学習させることによる認識精度向上

Transformer-based OCR(TrOCR)

Transformerベースのエンコーダ構造を持つ手書き文字認識モデル
→画像内の**広い範囲**との関係性を取り込んだ特徴の獲得が可能

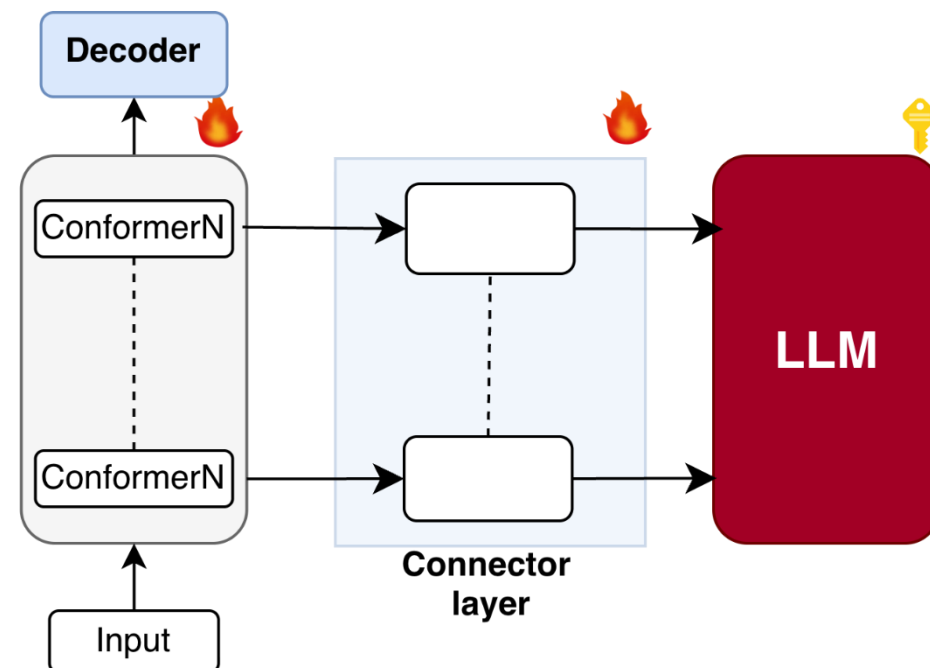


LLM損失によるエンコーダ学習

自動音声認識におけるLLM(Large Language Model)の活用

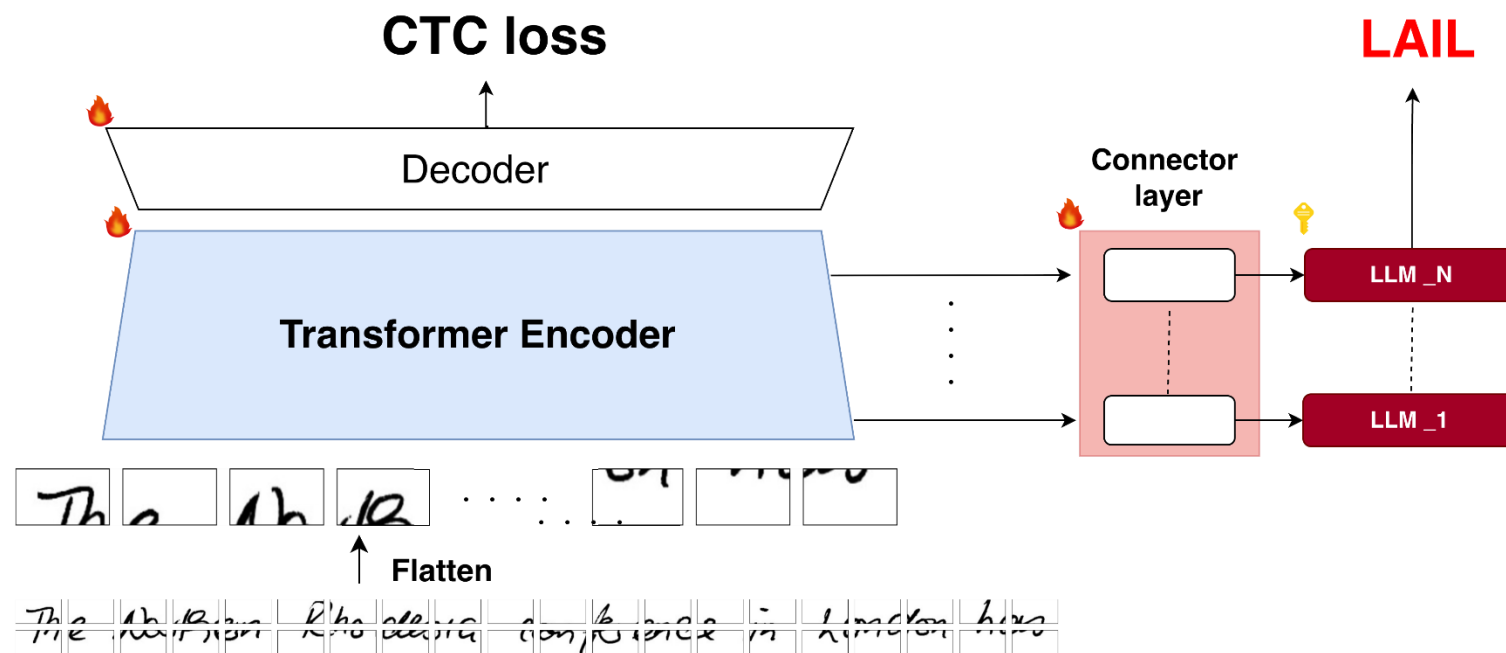
1. 複数のエンコーダ内の層からコネクタ層へ入力
2. 各コネクタ層からLLMへ
3. LLMから損失値を導出

→エンコーダが**言語知識**を獲得



エンコーダへの言語特徴付与 Model Architecture

エンコーダに**視覚特徴**・**言語特徴**の付与



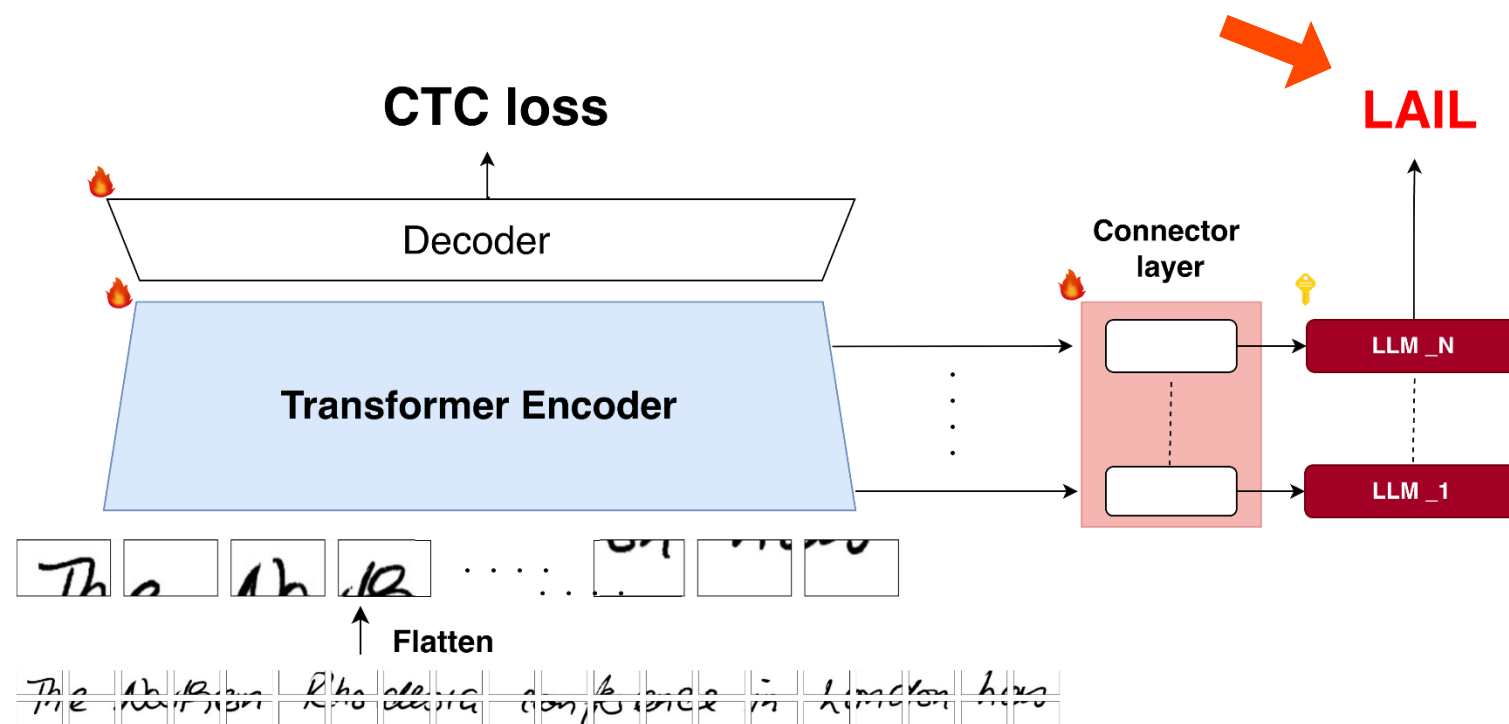
→CTC損失と**LAIL**損失を併用

エンコーダへの言語特徴付与 損失関数の設計

- LAIL 損失

$$\mathcal{L}_{LAIL} = - \sum_{l \in L} \lambda_l \sum_{t=1}^T \log P(\mathbf{y}_t | \mathbf{y}_{<t}, \mathbf{h}_{proj})$$

\mathbf{y}_t : 正解文字列
 $\mathbf{y}_{<t}$: tまでの正解文字列
 \mathbf{h} : コネクタ層の出力
 L : LAILへの接続層
 λ_l : 各層の損失の調整HP

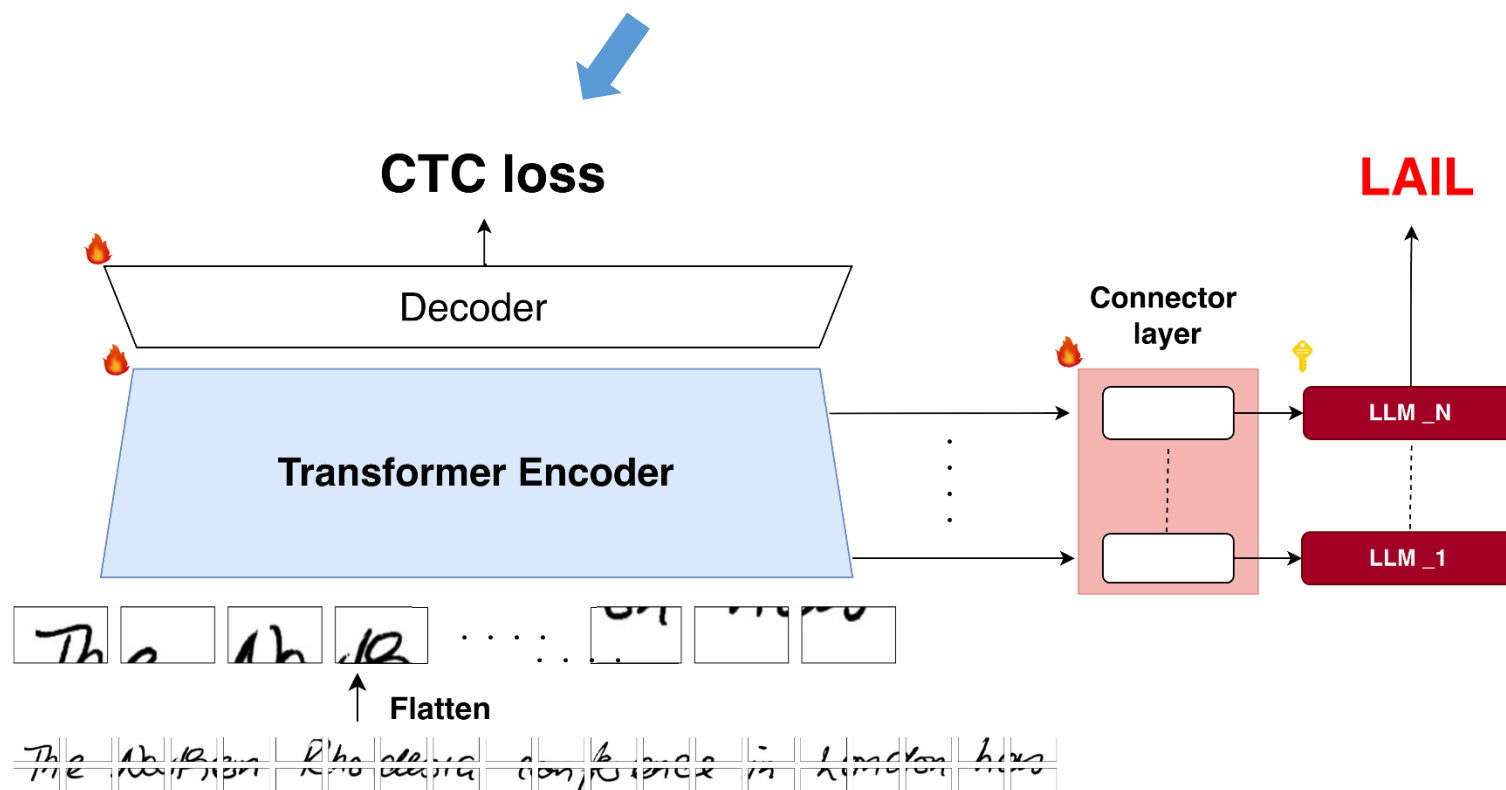


エンコーダへの言語特徴付与 損失関数の設計

- CTC 損失

$$\mathcal{L}_{CTC} = -\log(p(\mathbf{y}))$$

\mathbf{y} : 正解文字列
 $p(\mathbf{y})$: 予測文字列 \mathbf{y} になる確率



エンコーダへの言語特徴付与 損失関数の設計

- LAIL 損失

$$\mathcal{L}_{LAIL} = - \sum_{l \in L} \lambda_l \sum_{t=1}^T \log P(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{h}_{proj})$$

\mathbf{y}_t : 正解文字列
 $\mathbf{y}_{<t}$: tまでの正解文字列
 \mathbf{h} : コネクタ層の出力
 L : LAILへの接続層
 λ_l : 各層の損失の調整HP

- CTC 損失

$$\mathcal{L}_{CTC} = -\log(p(\mathbf{y}))$$

\mathbf{y} : 正解文字列
 $p(\mathbf{y})$: 予測文字列 \mathbf{y} になる確率

- 全体損失

$$\mathcal{L}_{total} = \mathcal{L}_{CTC} + \alpha \mathcal{L}_{LAIL} \quad \alpha : \mathcal{L}_{LAIL} \text{を調整するHP}$$

実験

実験設定

- IAMデータセット

- 画像高さ : 128px
- 学習データ数 : 6161枚
- 検証データ数 : 966枚
- テストデータ数 : 2915枚

- 実験設定(ベースライン)

- バッチサイズ : 4
- 学習率 : 1×10^{-4}

- 実験設定(提案手法)

- バッチサイズ : 4
- 学習率 : 1×10^{-4}
- 使用LLM : Llama3
- L_{LAIL} の重み α : 0.01

データセットについて

- IAMデータセット
 - 文字クラス数：79文字
 - 画像サイズ
 - 高さ：384px
 - 幅：2048px
 - 657 名の筆者によって記述
 - Aachen split
 - 同一筆者がテストデータと検証データ両方に入らないように分割
 - 筆者依存の過学習を避けた汎化性能を評価可能

実験

実験結果

1. ベースラインとの比較

	Baseline	Our Method
CER(%)	13.64	10.77

2. 配置による比較

Layer	10,11,12	4,8,12	3,12	6,12	11,12	12
CER(%)	12.54	11.96	12.55	13.16	10.77	12.44

3. LLM性能差比較

Number of LLM Parameters	1B	3B
CER(%)	12.44	12.66

→LLMサイズが大きすぎるとエンコーダが学習しきれない

実験

実験結果

出力例による本提案手法の有効性確認

come what may, we are never alone when

Ground Truth : come what may, we are never alone when

Baseline : come what may, we are never obome wlren

Our Method : come what may, we are never alone when

maturity. One remembered that he was a lawyer

Ground Truth : maturity. One remembered that he was a lawyer

Baseline : mnaturrrty. Ore remembered tthat he wos a lowyer

Our Method : maturity. One remembered that he was a lawyer

実験

評価指標

$$\text{CER} = \frac{S+D+I}{N}$$

S : 置換数

D : 削除数

I : 挿入数

N : 正解文字列の文字数

まとめ

結論

手書き文字認識モデルへの言語特徴付与は

→ 有効

より効果的にするには以下が重要

- コネクタ層の配置
- コネクタ層の数

留意点

LLMサイズ・数が大きすぎるとエンコーダが学習しきれない

→モデルの表現能力を考慮