進捗報告

1　今週やったこと

- Optuna でモデルのハイパーパラメータを調整すること
- ChineseBERT の資料を調べること

2　Optuna によるモデルのハイパーパラメータの調整

Optuna [1] は, ハイパーパラメータの最適化を自動化するためのライブラリです.

現在使っている灯謎問題を解く Seq2Seq モデルは Valid Data による Overfitting 問題を解決するため, Optuna で Encoder と Decoder の Dropout を調整しました.

漢字のみの実験結果は図 1 のように示します.

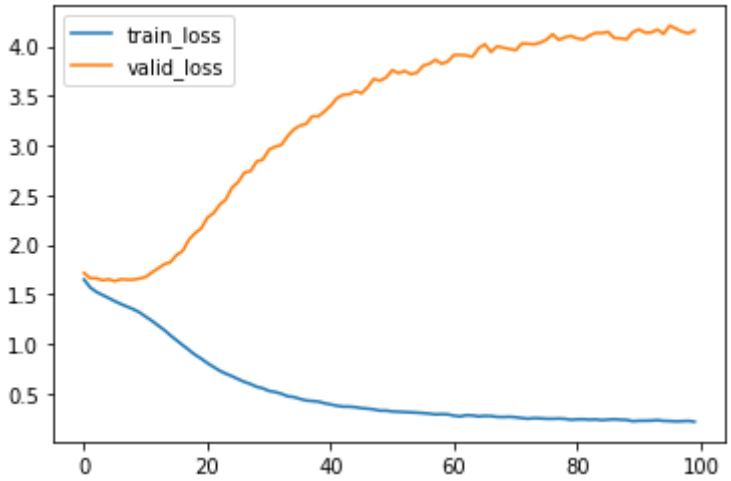

図 1: 漢字 (調整後) Train Loss と Validation Loss

Test Data の正解率は表 1 のように示します.

表 1: Test Data の結果

実験	Precision	Recall	F1
Char	0.576	0.771	0.631
Char(調整後)	0.611	0.750	0.644

100epoch の実験を経て Train Loss は 0.214 に, Valid Loss は 4.158 に収束しました.

結果として Train Loss は調整前の 0.688 より低いが, Valid Loss は調整前の 2.318 より高くなり, Overfitting 問題はより厳しくなりました.

Test Data について, 調整後の F1 値は 0.644 に上がりました.

他の条件の関する実験は進んでいますが, 出来次第レポートを更新します.

3　ChineseBERT に関する調査

ChineseBERT [2] は Sun らが提案した中国語専用のモデルです.

伝統の BERT モデルと比べ, ChineseBERT は漢字の形的情報 (Glyph embedding) と音声的情報 (Pinyin embedding) を考慮する特徴があります.

具体的には, 漢字を読み取る Char embedding の上, 漢字の形的情報を読み取る Glyph embedding と漢字の音声的情報を読み取る Pinyin embedding を導入することです. そしてこの 3 つの embedding を concatenate して, 全結合層で Fusion embedding を生成します. 最後 Fusion embedding と Position embedding を加算して BERT の入力とします.

図 2 に ChineseBERT の構造を示します.

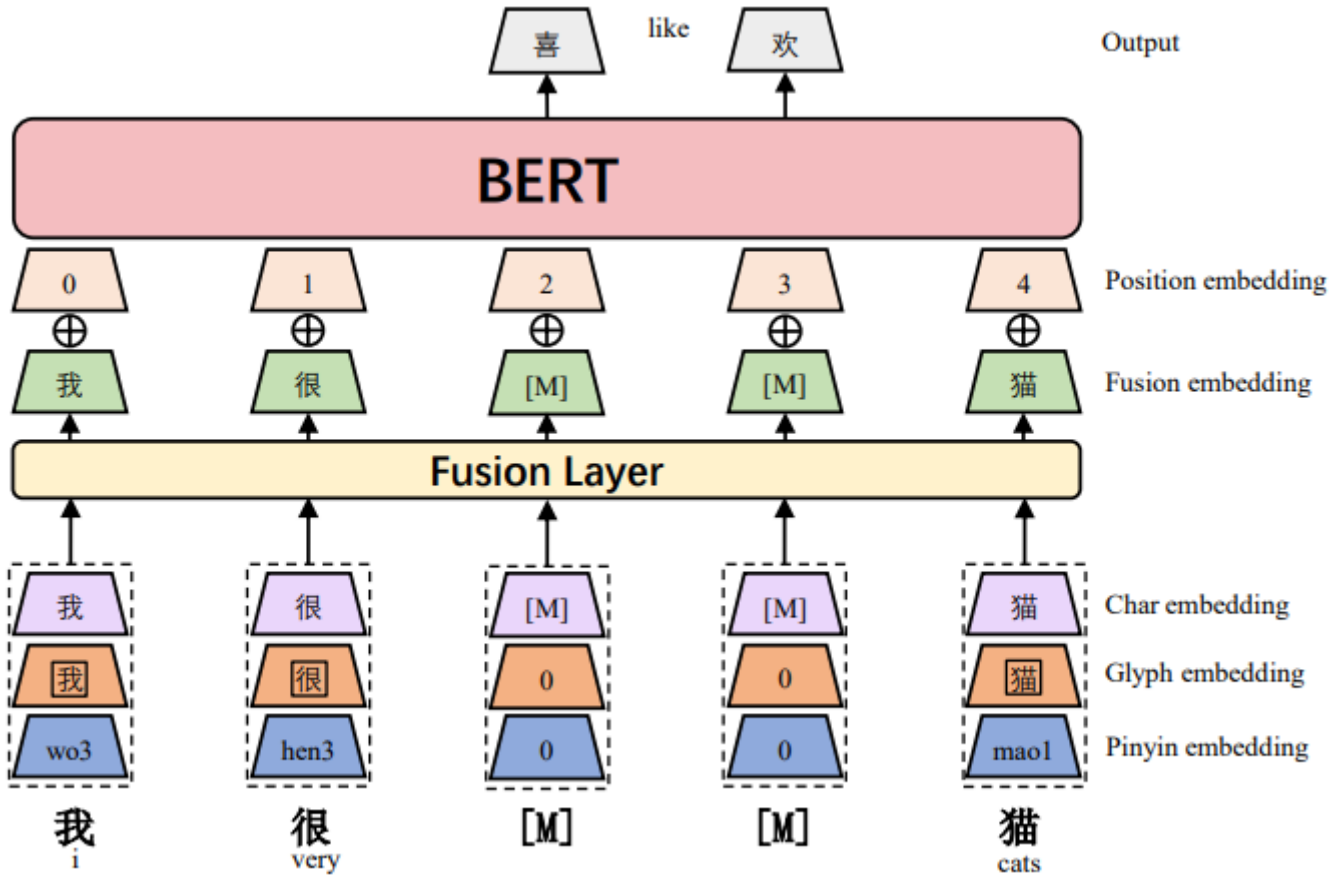

図 2: ChineseBERT [1] の構造

3.1　漢字の形的情報 (Glyph embedding)

Glyph embedding の構造について, まずは全ての漢字に対して, 三種類の font での 24×24 の画像を concatenate して $24 \times 24 \times 3$ の Tensor を生成します. そして全結合層で Glyph embedding を生成します.

図 3 に Glyph embedding の構造を示します.

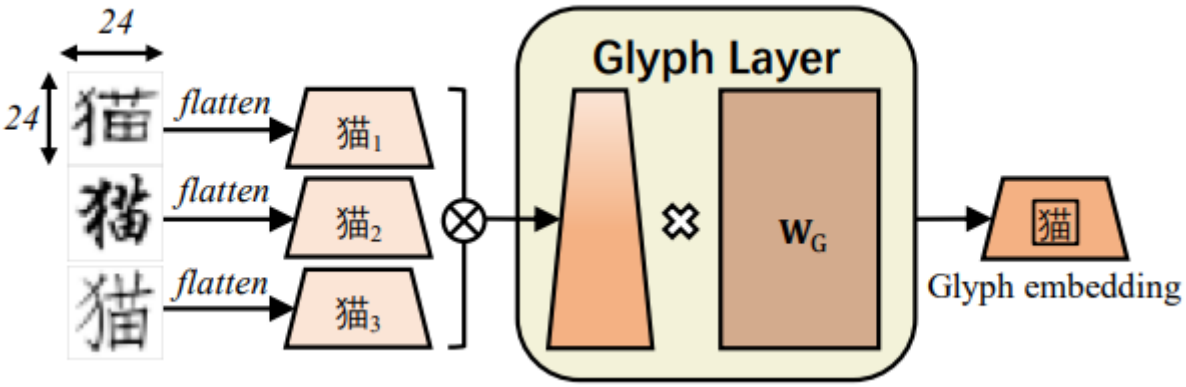

図 3: Glyph embedding の構造 [1]

3.2 漢字の音声的情報 (Pinyin embedding)

Pinyin embedding の構造について, まずは全ての漢字に対し pypinyin で漢字の Pinyin Sequence (アルファベットとイントネーション) を生成します. そして CNN で Pinyin embedding を生成します.
図 4 に Pinyin embedding の構造を示します.

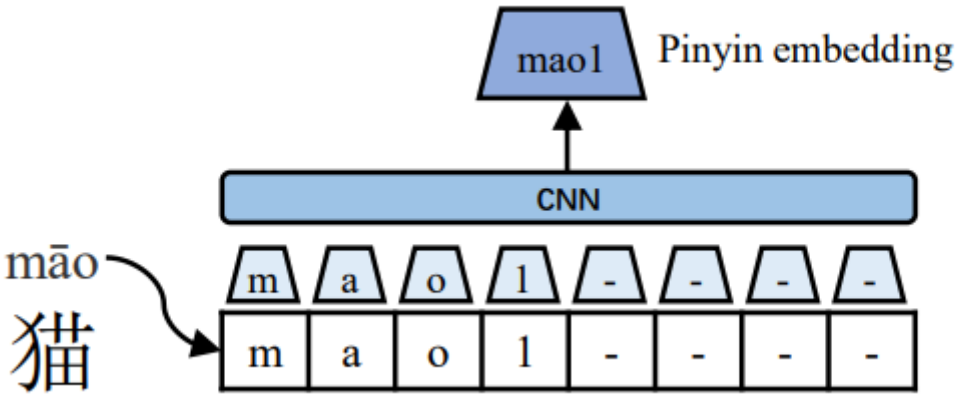

図 4: Pinyin embedding の構造 [1]

4 来週目標

- ChineseBERT を実装すること

参考文献

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, Vol. abs/1907.10902, , 2019.

[2] Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information, 2021.