令和 5 年 1 月 16 日 (水) ゼミ資料 M2 陳偉斉

進捗報告

1 今週やったこと

- データ削減 (1 対 1) での実験
- optuna による lstm パラメータの更新
- 「unk」トークンを削除した Bert モデルの実験

2 データ削減 (1 対 1) での実験

データ不均衡問題が実験結果に与える影響を確認するため, データセットを「正解, 不正解 (画が違う漢字), 不正解 (SUB 漢字が違う漢字)」から「正解, 不正解 (画と SUB 漢字が違う漢字各 50 %)」に変え, データ量 218811 (正解不正解 1 対 2) から 145874 (正解不正解 1 対 1) に変えました. そのため実験の結果も変わります.

3 実験内容

SVM, LSTM, bert-base-chinese モデルに対して実験しました. SVM(TfidfVectorizer) では「rbf kernel」を利用し,「正解不正解 1 対 1」と「正解不正解 1 対 1(ヒントなし)」二つの条件で実験を行いました. SVM(CountVectorizer) では「rbf kernel」を利用し,「正解不正解 1 対 1」と「正解不正解 1 対 1(ヒントなし)」二つの条件で実験を行っています (実験中). LSTM では「正解不正解 1 対 1」と「正解不正解 1 対 1(ヒントなし)」二つの条件で実験を行いました. BERT では「正解不正解 1 対 1」と「正解不正解 1 対 1(ヒントなし)」二つの条件で実験を行います (実験中).

4 実験結果

実験結果は表 1 に示します.

表 1: 各モデルの実験結果

結果	SVM(Count)	SVM(Count ヒントなし)	SVM(Tfidf)	SVM(Tfidf ヒントなし)	LSTM(27 epoch 現在)	LSTM(ヒントなし)	bert	bert(ヒントなし)
訓練誤差					0.11	0.22	0.09	未完成
訓練精度	0.60	未完成	0.59	0.60	0.95	0.89	未完成	未完成
テスト精度	0.10	未完成	0.10	0.09	0.89	0.84	0.74	未完成

完成次第表を補完します.

結果として, データの均衡はモデルに影響があります (前回は 0.66 と 0.56). 特に SVM に影響が大きいです. そして人類の判断に影響する「漢字の形」と「ヒント」が LSTM モデルに影響があります. 今回の実験は全部漢字を最小単位として扱うため, 画と SUB 漢字は次の実験に導入されてません. 次に導入します.

5 optuna によるパラメータの設定

optuna で 100 trials した結果で LSTM のパラメータを設定します. 表 2 に実験モデルのパラメータを示す.
bert は次にします.

表 2: 実験用パラメータ (LSTM)

パラメータ	数値
分散表現の次元数	464
隠れ層の次元数	334
隠れ層数	2
バッチサイズ	128
Dropout	0.06747612089075827
最適化手法	Adam
学習率	0.0032261837951104255
Epoch	100

6 来週目標

- 画と SUB 漢字を導入する (データの中に分けるか, 分散表現を生成するか)