令和 4 年 03 月 16 日 (水) ゼミ資料 M1 陳偉斉

進捗報告

1 今週やったこと

- 中国語 Word Vectors での対照実験

2 データセット

中華灯謎データベースから答えが一文字の灯謎 79727 問を収集した.

そして元々の問題文-正解ペアに対し, 同じ問題文-不正解のペアを作成した. 故に実験用データは 79727 問から 159454 問に拡張した.

これらのデータセットに正解は 1 , 不正解は 0 のようにラベルを付け, 灯謎の答えが正解か不正解かを判明するように実験した.

3 Chinese Word Vectors

Chinese Word Vectors[1] とは, 北京師範大学と人民大学の Shen によって公開した中国語 Pretrain 単語ベクトル表現のコーパスである.

このコーパスは, 十種類以上のデータセットと訓練手法のコンビネーションで訓練した単語ベクトル表現が含まれている.

今回は Wikipedia zh と Skip-Gram で訓練した単語の埋め込み表現を利用する.

4 実験用モデル

今回の実験は RNN と LSTM モデルを使用した.

対象実験として,RNN (漢字だけの分かち書き),RNN (単語だけの分かち書き),LSTM (漢字だけの分かち書き),LSTM (単語だけの分かち書き),LSTM (Wikipedia zh) 5 つの方法で 200 epoch の実験を行った.

5 実験結果

実験結果として,RNN に対し, 漢字分けの Train Loss と Validation Loss は 0.693 に収束し,Train Accuracy と Validation Accuracy は各自 50.05 と 49.92 パーセントに収束した. そして単語分けの Train Loss と Validation Loss も 0.693 に収束し,Train Accuracy と Validation Accuracy は各自 50.28 と 49.94 パーセントに収束した.

一方,LSTM に対し, 漢字分けの Train Loss と Validation Loss は各自 0.472 と 0.784 に収束し,Train Accuracy と Validation Accuracy は各自 75.87 と 63.89 パーセントに収束した. そして単語分けの Train Loss と Validation Loss 各自 0.305 と 0.985 に収束し,Train Accuracy と Validation Accuracy は各自 86.05 と 62.34 パーセントに収束した.

その他,LSTM に対し,Wikipedia zh の Pretrain ベクトルを利用した実験の Train Loss と Validation Loss は各自 0.078 と 1.309 に収束し,Train Accuracy と Validation Accuracy は各自 97.07 と 73.00 パーセントに収束した.

実験結果は図 1 , 図 2 , 図 3 , 図 4 , 図 5 のように示す.

結論として,RNN と比べると,LSTM の方が訓練データセットとバリデーションデータに対する表現がよいが, テストアキュラシィは各自 50 パーセント前後になるので, 他のモデルで実験する必要性がある.

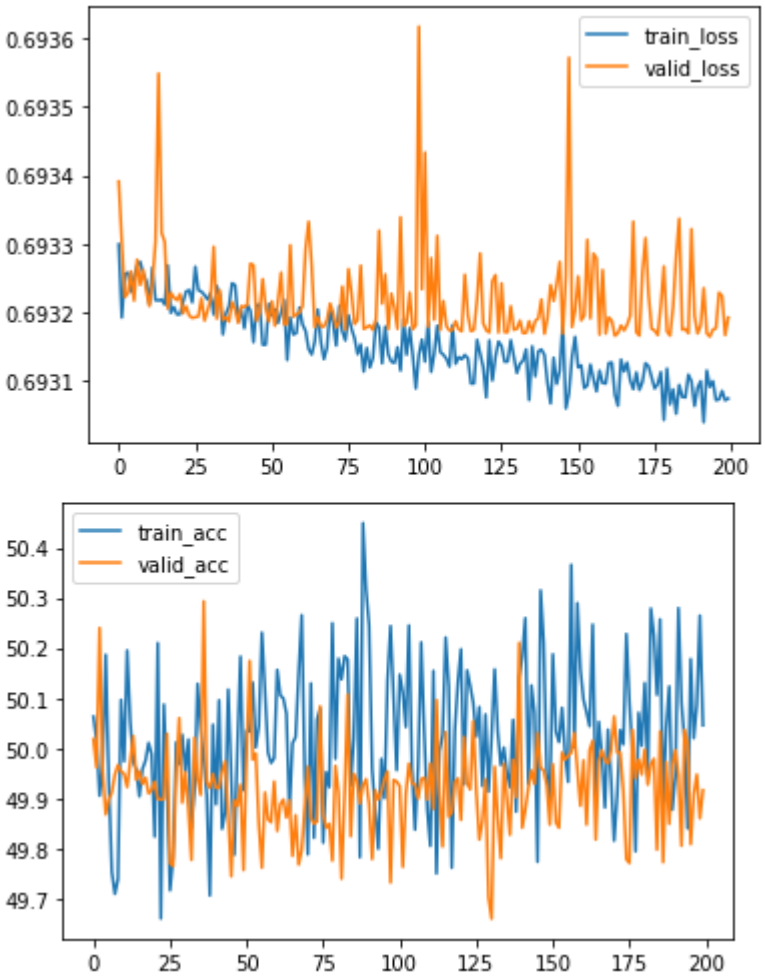
train_loss
valid_loss
0.6936
0.6935
0.6934
0.6933
0.6932
0.6931
0 25 50 75 100 125 150 175 200
train_acc
valid_acc
50.4
50.3
50.2
50.1
50.0
49.9
49.8
49.7
0 25 50 75 100 125 150 175 200

図 1: RNN 漢字

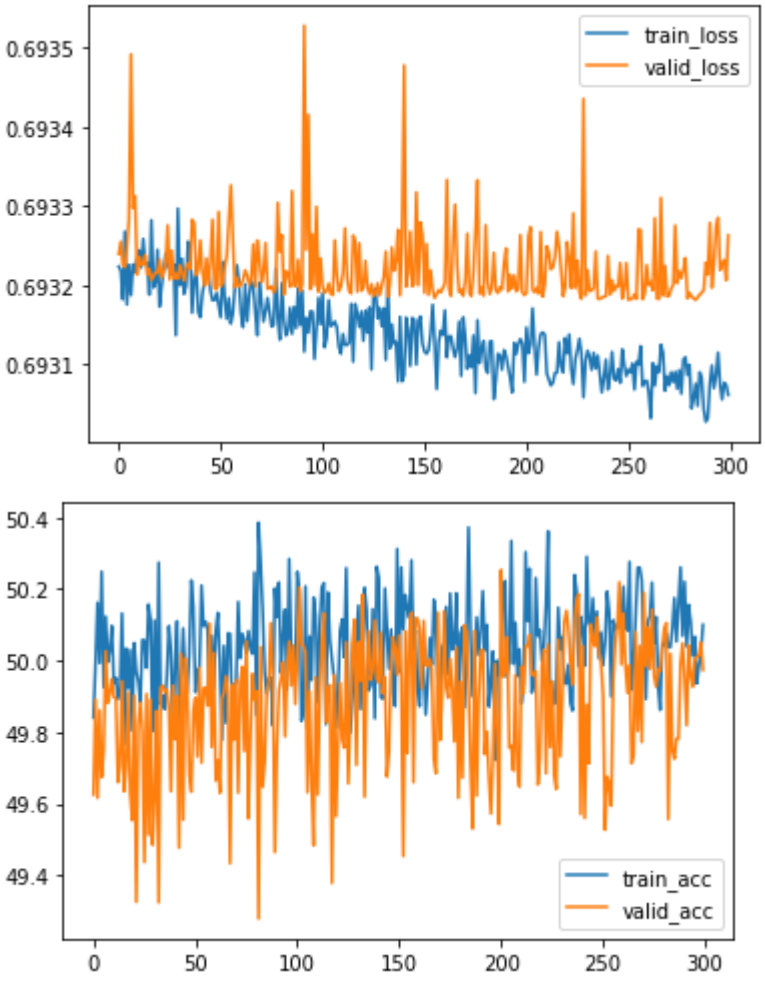
train_loss
valid_loss
0.6935
0.6934
0.6933
0.6932
0.6931
0
50
100
150
200
250
300
50.4
50.2
50.0
49.8
49.6
49.4
train_acc
valid_acc
0
50
100
150
200
250
300

図 2: RNN 単語

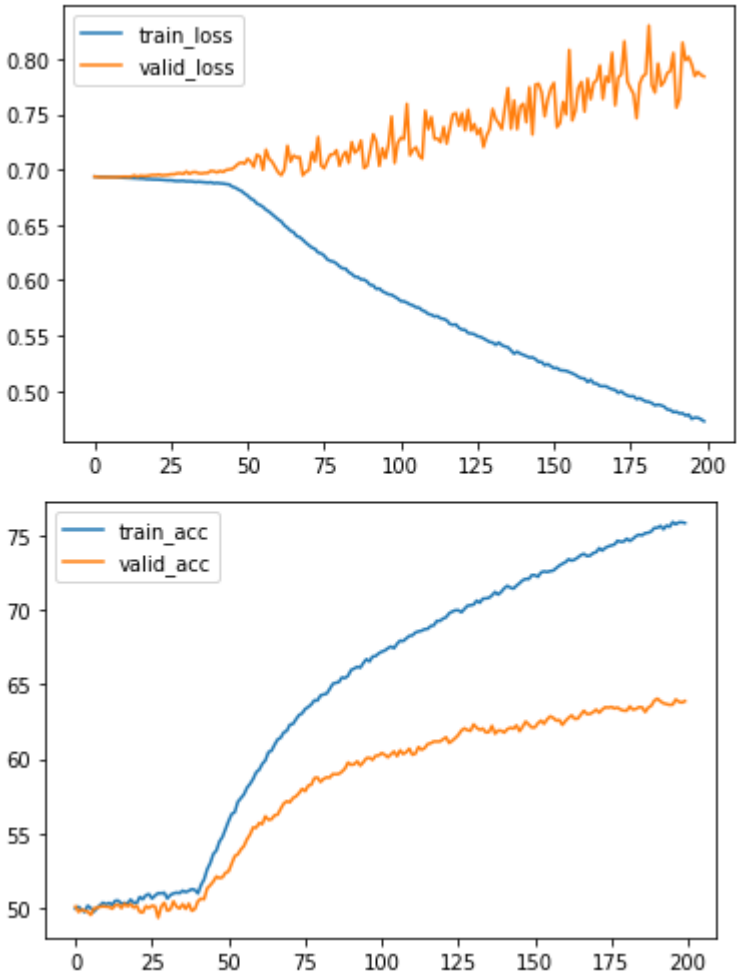
train_loss
valid_loss
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0
25
50
75
100
125
150
175
200
train_acc
valid_acc
75
70
65
60
55
50
0
25
50
75
100
125
150
175
200

図 3: LSTM 漢字

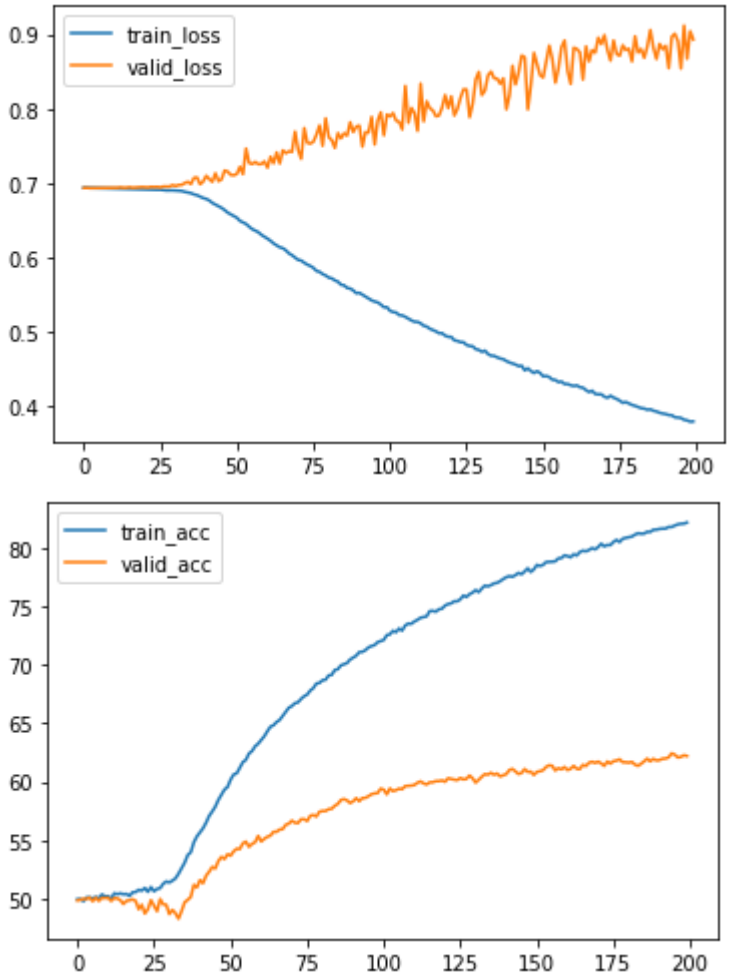
train_loss
valid_loss
0.9
0.8
0.7
0.6
0.5
0.4
0
25
50
75
100
125
150
175
200
train_acc
valid_acc
80
75
70
65
60
55
50
0
25
50
75
100
125
150
175
200

図 4: LSTM 単語

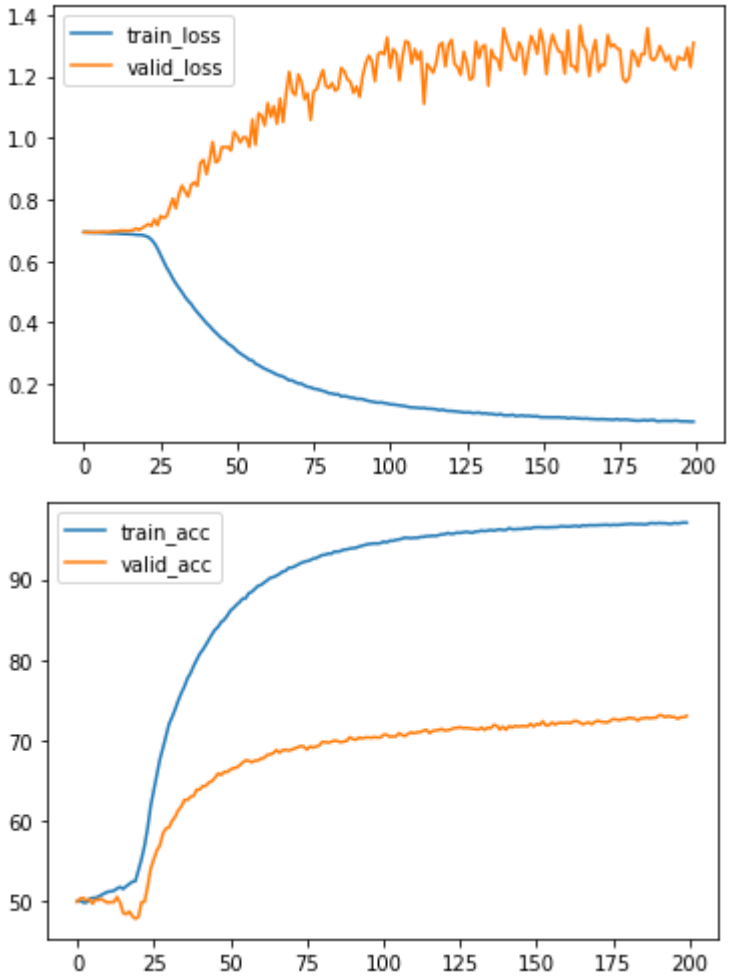

図 5: LSTMWikipedia

6　来週目標

- 他のモデルと word vector で実験すること
- 漢字成分情報の導入について検討すること
- 問題と答えの関係について検討すること

参考文献

[1] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 138–143. Association for Computational Linguistics, 2018.