令和 4 年 08 月 29 日 (月) ゼミ資料 M2 陳偉斉

進捗報告

1 今週やったこと

- ChineseBERT で灯謎から漢字の画像を生成する実験を行います (画像データ補足中).

2 ChineseBERT で灯謎から漢字の画像を生成する実験

ChineseBERT は Sun らが提案した中国語専用のモデルです. ChineseBERT は漢字の形的情報 (Glyph embedding) と音声的情報 (Pinyin embedding) を考慮したモデルであるため,SUB 漢字や画など漢字成分と違い効果を期待しています. 故に ChineseBERT を利用し, 灯謎を解答する可能性を確認します.

具体的には問題を ChineseBERT に入力し,ChineseBERT で生成した CLS 出力と全結合層で解析度 24×24 の漢字画像を生成します.

漢字分類モデルと比べると, 漢字の画像生成モデルは, 出力の次元数による削減 (8266 から 576) と Out-of-vocabulary (OOV) 問題による解消この二つの利点があります.

その一方, 漢字の画像生成には精確率の計算という問題があります.

この問題を解決するためまだ漢字になれるかどうかを確認する分類モデル 1 と, 出力した漢字は正解かどうかを確認する分類モデル 2 が必要となります. この部分は次の課題になります.

2.1 モデル

モデルについて, 今回は ChineseBERT-base を利用するため, 出力の次元数は 768 です.

出力層は 768×576 の全結合層で実現します.

2.2 データセット

「国標一級漢字 (3755 字)」で作成した「漢字」と「画像」は解析度 64×64 の黒白画像ですので, 漢字不足と画像解析度変換の問題があります.

現在漢字データの補足を進んでいます.

2.3 課題

- 漢字データの補足.
- 正解率を確認するモデルの構築.

3 来週目標

- 漢字のデータを増加し訓練します (灯謎に使われている 8000 字).