進捗報告

1 今週やったこと

- BERT の利用方法について慣れる事を目的に Google Colab で幾つかのサイトを参考にプログラムを動かした．そのうちのひとつの実験について結果を述べる．

2 実験の詳細

本実験は，「Google colab で BERT を使ってライブドアニュースコーパスを多クラス分類をする」 [*1] のサイトのプログラムを利用した．環境は Google Colab , モデルは BERT , 入力は記事のタイトル，出力は記事のジャンル (9 個) である．使ったデータセットはライブドアニュースコーパス [*2]である．前処理を施した後のデータサイズは 7325 行 3 列 であり，メディアのジャンルとラベルと記事のタイトルの 3 項目の列である．トークナイザーは東北大学学習済みモデル (cl-tohoku/bert-base-japanese-whole-word-masking) のトークナイザを使用．本実験の学習は，この東北大学習済みモデルをファインチューニングしてライブドアデータセットを学習させた．BERT は一つの文章のトークン数の上限が 512 であるので，それを超える文章があった場合は文章の長さをカットしなければいけないがこのデータセットの最大トークン数は 74+2 なので，そのまま使う．最適化手法には AdamW を使用した．ハイパーパラメータは，バッチサイズ 32, 学習率 1e-5, エポック数 4 で 学習させた．学習の際，訓練データはデータセットの 9 割を使用し，テストデータはその残りを使用した．学習させた結果，テストデータの正解率は 0.81 となった．図 1 に，Epoch 数と Loss との関係を示す．よって，4 以下の範囲内ならエポック数を増やすと正解値と予測値のずれが小さくなることが分かった．

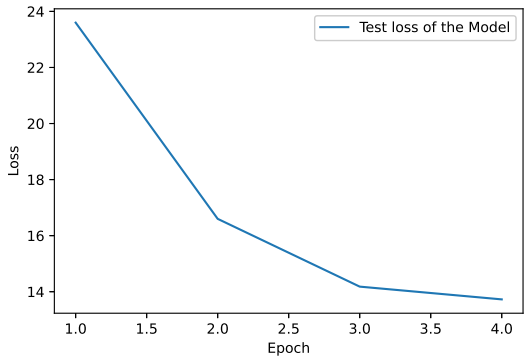

図 1 Test loss of the Model

3 まとめ

今回は BERT の習熟を目的としてライブドアニュースコーパスを 9 クラス分類した．今後はパラメータを動かすとどうなるか，と他の学習モデルとの精度を比較してみたいと思う．また，使えるデータセットを探したいと思う．

[*1] https://tech.fusic.co.jp/posts/2021-04-23-bert-multi-classification/
[*2] https://www.rondhuit.com/download.html#ldcc