CSC2515 Final Project - Part I

Building a Recommender System

Yuzhou Liu, 1007024836

Kaggle Username: yuzhouliu06

1. Introduction

Recommender systems are algorithms that could combine users and products. They aim to predict users' preference. Rating prediction is such an important task for a recommender system. It will predict a user's ratings for those items which were not rated yet by the user. Common methods of rating prediction are biased matrix factorization, collaborate filtering, etc. (Pero & Horváh, 2013)

The goal of this project is to perform a rating prediction task based on Amazon music reviews dataset. After probing through the dataset, I decide to use the review texts as my training data and train a BERT model through fine-tuning. I have achieved the optimal parameters within the computational capacity. The model is constructed within the framework of PyTorch and Transformers, and computation has been sped up by cloud GPU on Google Colab. The model ends up with a score of 0.51759 on Kaggle. Further improvements to my model could be considering larger sequence size, incorporating the information of time, or combining the review text with the rating scores. The detailed analysis is as follow.

2. Exploratory Data Analysis

2.1 Data Preparation

The size of the training dataset is 200,000 with 72,285 different reviewers and 25,493 different

items. We can notice that the dataset is very sparse. The following figures represent the univariate distribution of rating scores and review text length, and also the corresponding bivariate distribution. Most of the rating scores are positive with a global average score around 4. For the review text length, the average value is 726 characters with a wide range between 1 and 30, 370. For the regime of high rating score (rating = 5), the text length is more various with large extrema.

Figure 1. The Distribution of Rating Scores

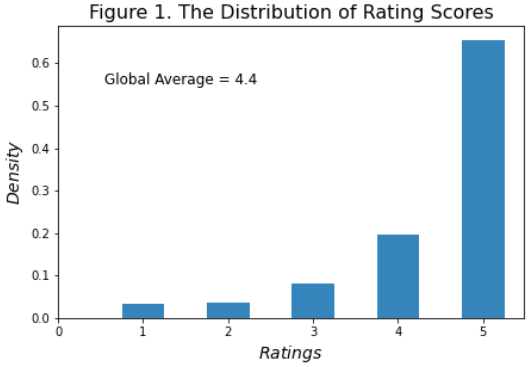


Figure 2. The Distribution of Review Text Length

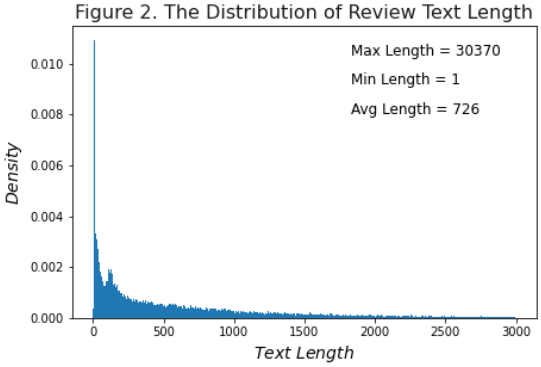
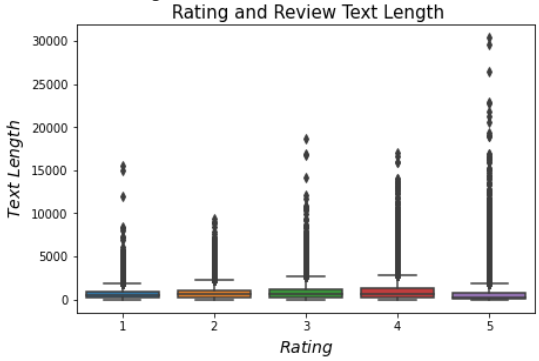


Figure 3. The Bivariate Distribution of Rating and Review Text Length



I decide to use review text to predict ratings considering users' attitudes towards the items are largely incorporated in their reviews. 20% of the training set are split for validation in each iteration. Also, the rows without review text have been removed from the training and predicting. During the final stage, the global average is used to make up for the missing score.

3. Modelling

3.1 Bidirectional Encoder Representations from Transformers

The BERT product rating predictor is a natural language processing model based on the

Bidirectional Encoder Representations from Transformers (BERT). It makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Transformer encoder reads the entire sequence of words at once.

Generally, a BERT-base model contains an encoder with 12 Transformer blocks, 12 self-attention heads, and the hidden size of 768. The input sequence of BERT is no more than 512 tokens. (Vasvani et al., 2017) [CLS] is added to head of sequence to embed class and [SEP] is used for separating sequences. For text classification tasks, a simple softmax classifier is added to the top of BERT to predict the label. During the training, the negative log likelihood loss and Adam algorithm is used for the multi-classification.

3.2 Results

In the first stage, I tried to train the model using all the dataset. The sequence length limit is 100. The batch size is 64. Results show that serious overfitting problem occurs when epoch number is greater than 3. Due to the computational capacity, I further tried to train the model setting sequence length as 200.

The model failed to finish training when applied larger sequence length to the overall training set. Also, the computing time becomes also longer. Therefore, 20% of training set data is sampled as a new training set. Sequence lengths of 300 and 512 are further tested with batch size also lowered to 32 and 16 respectively.

The best model yields a final loss of 0.19749 on the training set, 0.31502 in the validation set,

and 0.51759 on 50% of test data. The corresponding parameters are sequence length of 200, batch size of 64, and epoch of 2.

4. Conclusion

In this project, a BERT model is trained based on the Amazon review text to predict users' rating to the product. The performance of the optimal model exceeds the strong baseline on Kaggle with the score of 0.51759. But since only 50% of the test data is calculated, the final results remain uncertain. Still, the model could have several improvements:

- Subject to hardware condition, many combination of parameters are difficult to be tested. Training a BERT-large is more seemingly to be a mission infeasible
- The factor of time is not included into the model
- The rating score of the reviewers could be considered to adjust predicted score

Reference

1. Pero, Š. & Horváth, T. (2013). Opinion-Driven Matrix Factorization for Rating Prediction. *User Modeling, Adaptation, and Personalization Lecture Notes in Computer Science*, 1-13.
2. Vasvani A. et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.