Written Assignment (M)
2447939L Yize Li

PART A

SUMMARY OF THE 1st PAPER (BERT):

This paper introduces a new way to build a language representation model based on Transformer Encoder and can be applied to many language processing tasks.

The two existing strategies for applying pre-trained language representations to downstream tasks, feature-based and fine-tuning, use unidirectional language models to learn general language representations. Based on that, BERT introduced a bidirectional way into the pre-training process. It reuses the framework of OpenAI GPT and modifies its pre-training part to improve its performance.

For the pre-training procedure, BERT uses the BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words). For Wikipedia, BERT extracts only the text passages and ignores lists, tables, and headers. And for each fine-tuning task, it used 2.5k to 392k samples to train.

BERT model emphasizes the importance of using masked language model(MLM) and next sentence prediction(NSP). The model tops eleven natural language processing tasks, including pushing the GLUE score to 80.5, MultiNLI accuracy to 86.7 and so on.BERT model proves that a very deep model can significantly improve the accuracy of handling NLP tasks, and this model can be pre-trained from the unlabeled dataset. And its unexpected setup made it a very powerful model.

SUMMARY OF THE 2nd PAPER (A Hierarchical Location Prediction Neural Network for Twitter User Geolocation):

This paper introduces a hierarchical location prediction neural network for Twitter users. The problem is that though internet service providers can directly obtain users' location information from IP address, such private information is not available for third-party contributors.

Several ways have been applied like rule-based geographical references recognizer, deep-neural-network based methods and even hierarchical tree of earth grids. But existing ways of building a hierarchical tree is too time-consuming. This paper's method overcomes the issue and only needs to be trained once.

The paper suggests using datasets like Twitter-US, Twitter World and WNUT. The model combines 7 features of twitter and the user, handle them separately in different ways. Use Bi-LSTM to extract location-specific features from these text embedding vectors. After combining all text representations, plus adding some transformer encoder layers, the probability for each country is computed after a field-level attention layer and the city- level location is constrained by following layer to be more accurate.

The full model of this paper can reach an accuracy of 72.7%, 68.4% and 57.6% on each dataset separately. The relative country error can reach a minimum of 0.23 with the increase of the weight of country-level supervision signal. It not only improves the prediction accuracy but also significantly reduces the mean error distance. In the future, propagating features from connected friends would provide much more information than just using network embedding vectors, and potential improvements could also be made using better graph-level classification frameworks.

EVALUATION

Personally, my favorite part about the first paper is the thinking which inspired from previous work. To think deeper and wider about that unidirectional method can let the model improve so much. But to my opinion, the computation resource it requires is too much to be considered as an effective model. For the 2nd paper, the model performs very well and gives me another perspective of the application of language processing, but whether the user likes to be predicted his location based on the text he posted online even there is no location information in it or not, I think that's the key factor which requires debate apart from the technology itself.

APPLICATION

Because the BERT model has an outstanding performance over so many NLP tasks, it definitely can help to improve tasks about language understanding, like text classification or Q&A system. And for the 2nd model, user geolocation is a key factor for many important applications such as earthquake detection, disaster management and health management.

REFLECTION AND SYNTHESIS

BERT model gave me a new perspective of learning language representation. I was quite confused about how to fit one model to different levels of text like sentence-level and token level when first learned this in class. And by training MLM and NSP at the same time, it solves this problem. For the 2nd paper, at first, I don't quite understand how to predict users' locations based on features like twitter text, but the mechanism this paper introduces shows me that you can combine so many features to train the model simultaneously. And the hierarchical classifier can improve the performance by keeping digging into the text.

PART B

RESULT & ANALYSIS

As the figure shows that the performance of BERT model is better than Logistic Regression model in every way. The only thing is that training BERT takes much more time than training LR model.

```
Evaluation for: BERT
Classifier 'BERT' has Acc=0.930 P=0.957 R=0.946 F1=0.952
              precision    recall  f1-score   support

           0      0.858     0.884     0.871      2680
           1      0.957     0.946     0.952      7320

    accuracy                          0.930     10000
   macro avg      0.907     0.915     0.911     10000
weighted avg      0.930     0.930     0.930     10000


Confusion matrix:
 [[2368  392]
 [ 312 6928]]
```

```
Evaluation for: LR
Classifier 'LR' has Acc=0.865 P=0.929 R=0.891 F1=0.910
              precision    recall  f1-score   support

           0      0.696     0.784     0.737      2409
           1      0.929     0.891     0.910      7591

    accuracy                          0.865     10000
   macro avg      0.812     0.838     0.823     10000
weighted avg      0.873     0.865     0.868     10000


Confusion matrix:
 [[1889  826]
 [ 520 6765]]
```

Fig 1 Evaluation Metrics of BERT and LR Model

REFLECTION

During the process of solving the problem, I quickly learned how to apply the BERT model to another dataset and the basic structure of the model. At first, dreading the paper is very tough but when I combined reading with implementing the model in Colab it becomes much clearer. The second paper also introduces me of the current progress in the NLP field.

The most challenging part is trying to understand what does the code mean, but the example of BERT model is quite clear and many of them is in embedded within some functions so I don't need to fully understand it, just to know how to use it is enough. Now I got a new and powerful tool to handle problems about language processing.