A WWMBERT-based Method for Improving Chinese Text Classification Task

Xinyuan Wang, Inwhee Joe*
Dept. of Software Engineering, Hanyang University

중국어 텍스트 분류 작업의 개선을 위한 WWMBERT 기반 방식

왕흠원, 조인휘*
한양대학교 컴퓨터소프트웨어학과
xinyuanwanggo@gmail.com, *iwjoe@hanyang.ac.kr

Abstract

In the NLP field, the pre-training model BERT launched by the Google team in 2018 has shown amazing results in various tasks in the NLP field. Subsequently, many variant models have been derived based on the original BERT, such as RoBERTa, ERNIEBERT and so on. In this paper, the WWMBERT (Whole Word Masking BERT) model suitable for Chinese text tasks was used as the baseline model of our experiment. The experiment is mainly for "Text-level Chinese text classification tasks" are improved, which mainly combines Tapt (Task-Adaptive Pretraining) and "Multi-Sample Dropout method" to improve the model, and compare the experimental results, experimental data sets and model scoring standards Both are consistent with the official WWMBERT model using Accuracy as the scoring standard. The official WWMBERT model uses the maximum and average values of multiple experimental results as the experimental scores. The development set was 97.70% (97.50%) on the "text-level Chinese text classification task". and 97.70% (97.50%) of the test set. After comparing the results of the experiments in this paper, the development set increased by 0.35% (0.5%) and the test set increased by 0.31% (0.48%). The original baseline model has been significantly improved.

1. Introduction

Today, pre-training models are already an important part of the NLP field. A good pre-training model can train good "semantic representations". For example, Chinese words are composed of words, and "character vectors" are used to represent Chinese characters. , So the pre-training model needs to train better "character vectors" to represent Chinese words. The "character vectors" are used as raw materials for downstream tasks. The higher the quality of the "character vectors" obtained by the pre-training model, the results obtained during the downstream task experiment The better, so the goal of this paper is to get a better pre-training model based on the original baseline pre-training model through method improvement, so as to improve the effect of downstream tasks.

The BERT pre-training model [1] was launched by the google team in 2018. It is also a representative pre-training model in the NLP field. The model uses the "Encoder" part in Transformer[2] and the Attention mechanism, which can better capture the semantic information of the context, plus the use of "Masked LM" and "Next Sentence" The two language model tasks "Prediction" learn the "semantic representations" at the "word" and "sentence" levels respectively, so as to learn better "word vectors". The effects in each NLP downstream task surpass other models.

The baseline model used in this experiment is the "WWMBERT" model for Chinese [3]. This model uses the Whole Word Mask method on the basis of the original BERT model, which makes the pre-training model more obvious in the task part of the MLM language model.

Recently, the unsupervised Tapt(Task-Adaptive Pretraining) method [5] has significantly improved the effect of the model pre-training phase. This experiment will also combine this method to improve the model.

In deep learning model training, we often use the Dropout method [4] to prevent the model from overfitting. Later, we also derived a method called "Multi-Sample Dropout" [6], which reduced the number of training iterations, and It can also reduce the error rate and loss of the training set and the validation set. In this paper, through the combination of Tapt (Task-Adaptive Pretraining) method [5] and Multi-Sample Dropout method [6] to improve the baseline model, the experimental results show that the combination of the two methods is feasible, and The model effect has been greatly improved.

2 Related work

2.1 WWMBERT

In BERT's MLM language model task, the original word segmentation method based on "WordPiece" will divide a complete word into several sub-words. When training samples are generated, these separated sub-words will be randomly covered. In "Whole Word Mask", if part of "WordPiece" in a complete word is masked, other parts that belong to the word will also be masked, that is, WWM (Whole Word Mask). In the same way, since Google officially released official BERT Chinese model(BERT-base,Chinese), Chinese is segmented at the granularity of characters, without considering the Chinese word segmentation in traditional NLP. Therefore, the "Whole Word Mask" method is applied to Chinese, using Chinese Wikipedia (including simplified and traditional) to train the

pre-training model, and using "Harbin Institute of Technology LTP" as a word segmentation tool. [3]

2.2 Tapt (Task-Adaptive Pretraining)

The method advocates the use of unlabeled task data sets related to downstream tasks of the pre-training model to continue unsupervised training of the pre-training model. The model is pre-adapted to the downstream task in advance, and then fine-tuned for the downstream task, and supervised training is performed to obtain better results. Experiments have found that this method is more suitable when there is a small amount of data. [5]

2.3 Multi-Sample Dropout

The traditional Dropout [4] method randomly selects a set of samples (Dropout samples) from the input during each round of training, while in Multi-Sample Dropout [6], multiple Dropout samples are created, and the losses of all samples are averaged. So as to get the final loss. This method only needs to copy part of the training network after the Dropout layer, and share the weights between these copied fully connected layers. By synthesizing the loss of M Dropout samples to update the network parameters, the final loss is more than any single "Dropout" Samples' losses are all low. The effect of this is similar to repeating training M times for each input in a Minibatch. Therefore, it greatly reduces the number of training iterations, and can also reduce the error rate and loss of the training set and the validation set.

3 Methodology

The experiment combines the Tapt (Task-Adaptive Pretraining) [5] and Multi-Sample Dropout [6] methods to improve the model.

3.1 pre-train

We hope that the pre-training model can be more compatible with the downstream task data set before performing downstream tasks. The Tapt method can use unsupervised learning to allow the trained pre-trained model to continue training on the downstream task-related data set when it has a small amount of data and the training data it holds is in the same field as the downstream task data set, So that the pre-training model is more "familiar" with the upcoming downstream task data.

3.2 Fine-tuning

In the Fine-tune stage, the downstream task training set data is used to perform labeled supervised learning of the model. Due to the large amount of data, the model training time will be very long, and the phenomenon of "overfitting" is prone to occur, so during the training Applying the "Multi-Sample Dropout method" [6], the model can not only reduce the number of iterations to speed up the training speed, but also reduce the error rate of the model, improve the accuracy rate, and finally combine multiple parameter adjustment experiments, so that the model is in the original experimental effect The foundation has been greatly improved.

4 Experiments

4.1 DataSet

The data set used in the experiment is the same as the original official model. The news data set "THUCNews" released by the Natural Language Processing Laboratory of Tsinghua University is used. News needs to be divided into one of 10 categories, which is a 10-category chapter collection text Classification tasks. The evaluation index is: Accuracy

4.2 Experiment of settings

In the pre-training part, the baseline model we used is Chinese WWMBERT[3]. The basic configuration of the model hyperparameters is consistent with the original official BERT basic configuration, including 12 "transoformer blocks" and 110m "parameters", "Attention to multiple heads" The number of "forces" is 16, and the output hidden layer size is 768. On the basis of the original pre-training model, a subset of the THUNews data set was selected. After the data pre-processing of the subset, the Tapt method [5] was used to continue the unsupervised MLM (Masked Language Model) on the pre-training model. Task, after adapting the model to the task domain data set in advance, then perform the downstream text classification task.

In the Fine-tune part, the Multi-Sample Dropout method [6] is used for reference. The number of "Dropout Samples" in the original paper is 2, and the number of "Dropout Samples" used in this experiment is 5 after several attempts. Therefore, after each minibatch is trained, 5 different training losses will be obtained, and then the 5 different loss values will be summed and averaged to obtain the final loss, do backpropagation, and update the model parameters.

4.3 Experiments of results

Due to the huge data set, each round of the experiment takes about 10 hours in a 4*2080Ti GPU environment. Three experiments of "combined with TAPT", "combined with Multi-Sample Dropout", and "combined with TAPT and Multi-Sample Dropout" were carried out. Each experiment was carried out for three rounds, and the final "experiment result" used the "maximum value (Average value)" format is consistent with the official evaluation standard format of the baseline model. The following table:

TAPT: Task-Adaptive Pretraining
MD: Multi-Sample Dropout

Model/Accuracy	Dev (acc%)	Test (acc%)
BERT-wwm (baseline model)	97.70(97.50)	97.70(97.50)
+TAPT	97.82(97.72)	97.78(97.57)
+MD	97.96(97.86)	97.73(97.60)
+TAPT+MD	**98.05(98.00)**	**98.01(97.98)**

Table 1. Model comparison results

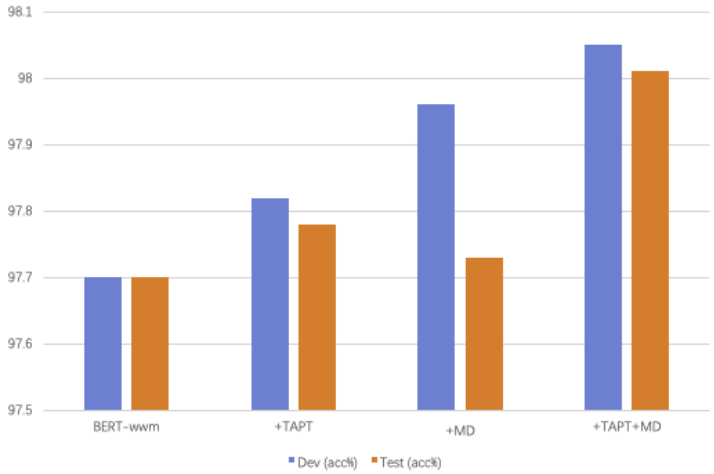

Figure 1 Model comparison results

It can be seen from the experimental results that using the "TAPT" method [5] or the "Multi-Sample Dropout" method [6] alone does not improve the model significantly, and even reduces the model performance. However, combining the two methods with appropriate parameter adjustments, the results obtained greatly surpass the original baseline model effect. The results show that the development set has increased by 0.35% and the test set has increased by 0.31%. This shows that our combination of methods and adjustment of model parameters have had a significant effect on the model.

5 Conclusion

In this paper, I propose a method to combine the Tapt method with the Multi-Sample Dropout method, and modify and experiment the number of Dropout Samples in the latter method. Compared with the baseline model, it does not expand the parameters, keep Based on the original model configuration, the model has been greatly improved and improved. Due to the limited hardware equipment of this experiment, coupled with the size of the data set and model, each round of the experiment requires 10 hours, and the number of parameter adjustments is also limited. It may not achieve the best results. From the current point of view, the relative Compared with the official "WWMBERT" model, the effect has been greatly improved.

Reference

[1] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need[J]. arXiv, 2017.

[3] Cui Y, Che W, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT. 2019.

[4] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4):págs. 212-223.

[5] Gururangan S, A Marasović, Swayamdipta S, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks[J]. 2020.

[6] Inoue H. Multi-Sample Dropout for Accelerated Training and Better Generalization[J]. 2019.