MiniRBT: A Two-stage Distilled Small Chinese Pre-trained Model

Xin Yao†, Ziqing Yang†, Yiming Cui‡†, Shijin Wang†
†State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Beijing, China
‡Research Center for SCIR, Harbin Institute of Technology, Harbin, China
†{xinyaol0,zqyang5,ymcui,sjwang3}@iflytek.com
‡ymcui@ir.hit.edu.cn

Abstract

In natural language processing, pre-trained language models have become essential infrastructures. However, these models often suffer from issues such as large size, long inference time, and challenging deployment. Moreover, most mainstream pre-trained models focus on English, and there are insufficient studies on small Chinese pre-trained models. In this paper, we introduce MiniRBT, a small Chinese pre-trained model that aims to advance research in Chinese natural language processing. MiniRBT employs a narrow and deep student model and incorporates whole word masking and two-stage distillation during pre-training to make it well-suited for most downstream tasks. Our experiments on machine reading comprehension and text classification tasks reveal that MiniRBT achieves 94% performance relative to RoBERTa, while providing a 6.8x speedup, demonstrating its effectiveness and efficiency.

1 Introduction

In recent years, the pre-trained language model based on Transformers (Vaswani et al., 2017) has become a paradigm of natural language processing, and their performance has been increasing with the growth in model size. These models have dominated the lists of primary AI models for natural language processing and computer vision, including BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019), SpanBERT (Joshi et al., 2020), ELECTRA (Clark et al., 2020). Despite their significant progress, pre-trained models still face considerable challenges in practical applications, such as high training costs and high latencies. Furthermore, while research on pre-trained models has mainly focused on the English language, there is a lack of smaller pre-trained models for Chinese.

Therefore, to further advance the development of Chinese natural language processing, we propose a small Chinese pre-trained model with strong practicability. Our approach involves utilizing dynamic whole word masking techniques (Cui et al., 2021) to generate training samples that facilitate comprehensive modeling of coarse-grained semantics. We apply a narrow and deep model structure for the student model. During the pre-training, we employ a two-stage distillation method that incorporates a teacher assistant model. We first distill from the teacher to the teacher assistant, and then from the teacher assistant to the student. This results in a lighter and faster model that can be fine-tuned for multiple downstream tasks, demonstrating excellent performance.

Our research focuses primarily on developing small Chinese pre-trained models that can be applied to practical tasks. The main contributions of this paper can be summarized as follows:

1. We conduct experiments that conclude that a narrower and deeper network structure is more effective than a wide and shallow structure of similar size.
2. Based on the above finding, we propose MiniRBT†, a small pre-trained model for Chinese which is only 10% the size of Chinese RoBERTa, while maintaining an average performance of 94% compared to RoBERTa. Moreover, it offers a 6x-7x speedup.

2 Background

2.1 Chinese RoBERTa-wwm

Since the emergence of BERT, pre-trained language models have advanced significantly and rapidly. However, conventional Chinese pre-trained models, such as the original Chinese BERT, typically employ a segmentation method that divides Chinese sequences at the granularity of individual characters, which means that Chinese words

†Available at https://github.com/iflytek/MiniRBT
Table 1: Examples of different masking strategies.

<table>
<thead>
<tr>
<th>Masking Strategy</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original text</td>
<td>使用语言模型来预测下一个词的概率。</td>
</tr>
<tr>
<td>Word segmentation</td>
<td>使用语言模型来预测下一个词的概率。</td>
</tr>
<tr>
<td>Original masking</td>
<td>使用语言[MASK]型来预测下一个词的概率。</td>
</tr>
<tr>
<td>WWM</td>
<td>使用语言[MASK][MASK][MASK]来预测下一个词的概率。</td>
</tr>
</tbody>
</table>

Figure 1: Comparison of one-stage distillation (left) and two-stage distillation (right) processes.

are segmented into characters using a WordPiece-based word segmentation approach. These characters are randomly masked individually during training sample generation without considering Chinese word segmentation. To address this issue, Chinese RoBERTa-wwm (Cui et al., 2021) uses a whole word masking (WWM) method specifically designed for Chinese. When part of a whole Chinese word is masked, other parts of the same word are also masked. It should be noted that the WWM method only affects the selection of masking tokens during the pre-training stage. Chinese RoBERTa-wwm is trained on Chinese Wikipedia (both Simplified and Traditional). In our work, we utilize Chinese RoBERTa-wwm as our teacher model.

2.2 Knowledge Distillation

In recent years, a growing number of works on model compression have been proposed to reduce the number of model parameters and improve the speed of model inference, such as quantization (Shen et al., 2020; Fan et al., 2020), pruning (Xia et al., 2022; Lagunas et al., 2021), knowledge distillation (Jiao et al., 2020; Sun et al., 2020; Sanh et al., 2019; Hou et al., 2020). Knowledge distillation (KD) transfers the knowledge embedded in a large teacher model to a small student model by mimicking the behaviors of the teacher. For instance, TinyBERT (Jiao et al., 2020) first obtains a general distilled small model by performing general distillation on the large-scale corpus from a general domain, and then performs task-specific distillation with downstream data during the fine-tuning stage. DynaBERT (Hou et al., 2020) trains a width-adaptive and depth-adaptive BERT by distilling knowledge from full-sized models to small sub-networks. KD is also applied to pruning to enhance performance, such as block pruning (Lagunas et al., 2021) and CoFi (Xia et al., 2022).

3 Method

In this section, we introduce the relevant methodologies employed by MiniRBT.

3.1 Whole Word Masking

The original WordPiece-based word segmentation method separates Chinese input sequences into independent characters and randomly masks the characters. This method simplifies prediction since the model can only remember particular character orders in words to predict masked characters without considering the semantic contextual relationships. To enhance the model’s performance, we adopted the dynamic Whole Word Masking (WWM) approach as a training sample generation strategy during the pre-training. We begin with utilizing the conventional Chinese word segmentation (CWS) method to segment the input into Chinese words, and are subsequently masked using WWM, as presented in Table 1. We employed LTP (Che et al., 2010) as our preferred tool to extract word
3.2 Two-stage Distillation

The traditional knowledge distillation method transfers knowledge directly from the teacher to the student. However, when there is a significant difference in the structures of the teacher and student models, this approach may result in a performance gap. To address this issue, we proposed using a two-stage distillation approach during the pre-training stage, which builds on the concept of Teacher Assistant Knowledge Distillation (Mirzadeh et al., 2020). As depicted in Figure 1, this method involves distilling knowledge from the teacher (RoBERTa) to the teacher assistant (RBT6), and then from the teacher assistant to the student (MiniRBT). The intermediate step of the teacher assistant helps to reduce the size gap between the teacher and the student model, subsequently improving the student models’ performance in downstream tasks.

To apply knowledge distillation (KD) with hidden layer distillation and prediction layer distillation, we employed TextBrewer (Yang et al., 2020), a PyTorch-based model distillation toolkit designed for natural language processing. We distill the knowledge from the output of the hidden layer. The objective is

\[ \mathcal{L}_{\text{layer}} = \sum \text{MSE}(H^t_i W^h_i, H^s_i) \]  

where the matrices \( H^s_i \in \mathbb{R}^{l \times d'} \) and \( H^t_i \in \mathbb{R}^{l \times d} \) represent the hidden representation of the \( i \)-th student’s hidden layer and the \( i \)-th teacher’s hidden layer respectively. The \( W^h_i \in \mathbb{R}^{d' \times d} \) is a linear transformation that matches the hidden state of the student network and the hidden state of the teacher network. Apart from mimicking the hidden layer behavior of the teacher, we also trained the student model by employing the cross-entropy loss with the teacher’s soft target probability

\[ \mathcal{L}_{\text{pred}} = -p(z^T) \cdot \log p(z^S) \]  

where \( z^S \) and \( z^T \) are the logits vectors predicted by the student and teacher respectively, and \( p = \text{softmax}(z/t) \) is the scaled probability with temperature \( t \) and logits \( z \).

Finally, we combine the hidden layer distillation with the prediction layer distillation:

\[ \mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{layer}} + \mathcal{L}_{\text{pred}} \]  

3.3 Narrower and Deeper Students

Through preliminary experiments, we find that a narrow and deep model structure outperforms a wide and shallow one, when they have the same number of parameters. Hence, we employed a narrow and deep design for MiniRBT. We present the details of the model structure in Table 2. MiniRBT consists of two branches of models, MiniRBT-H256 and MiniRBT-H288. These models have hidden layer dimensions of 256 and 288, respectively, and contain 6 transformer layers, pre-trained via the two-stage distillation approach.

4 Experiments

4.1 Downstream Tasks

Machine Reading Comprehension Machine reading comprehension (MRC) is a document-level modeling task that requires models to answer questions based on given passages. We evaluated our models on two Chinese reading comprehension datasets: CMRC 2018 (Cui et al., 2019) and DRCD (Shao et al., 2018). They are similar in the form of SQuAD (Rajpurkar et al., 2018). The evaluation metrics are F1 and EM.

Text Classification For single sentence classification, we use TNEWS and ChnSentiCorp (Tan and Zhang, 2008). The ChnSentiCorp dataset involves sentiment classification wherein texts need to be classified as either positive or negative, while TNEWS dataset involves the classification of short texts into various news categories. For sentence pair classification, we select three datasets: OCNLI, LCQMC (Liu et al., 2018), and BQ corpus (Chen et al., 2018). Both OCNLI and TNEWS are included as subtasks in the Chinese Language Understanding Evaluation (CLUE) Benchmark (Xu et al., 2020). The evaluation metric for these tasks is accuracy.

4.2 Training Setup

During the pre-training phase, a training batch size of 4096 and a peak learning rate of 4e-4 are employed, while the temperature is set to 8 and the number of training steps is 100K.

We fine-tune the model on downstream tasks for 2, 3, 5, and 10 epochs, respectively, with a learning rate chosen from \{5e − 5, 1e − 4\}. To decrease the impact of randomness on the experimental results, we run each task at least three times with different random seeds and report the average performance.
Table 2: Comparison of model structures. RBT stands for RoBERTa, and RBT3 is initialized by RoBERTa’s first three layers and pre-trained for 1M steps. The number of layers does not include the embedding and prediction layers.

<table>
<thead>
<tr>
<th>Model</th>
<th>Layers</th>
<th>Hidden Size</th>
<th>FFN Size</th>
<th>Heads</th>
<th>Model Size</th>
<th>Model Size (w/o embeddings)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa-wwm</td>
<td>12</td>
<td>768</td>
<td>3072</td>
<td>12</td>
<td>102.3M (100%)</td>
<td>85.7M (100%)</td>
<td>1x</td>
</tr>
<tr>
<td>RBT6 (KD)</td>
<td>6</td>
<td>768</td>
<td>3072</td>
<td>12</td>
<td>59.8M (58.4%)</td>
<td>43.1M (50.3%)</td>
<td>1.7x</td>
</tr>
<tr>
<td>RBT3</td>
<td>3</td>
<td>768</td>
<td>3072</td>
<td>12</td>
<td>38.5M (37.6%)</td>
<td>21.9M (25.6%)</td>
<td>2.8x</td>
</tr>
<tr>
<td>RBT4-H312</td>
<td>4</td>
<td>312</td>
<td>1200</td>
<td>12</td>
<td>11.4M (11.1%)</td>
<td>4.7M (5.5%)</td>
<td>6.8x</td>
</tr>
<tr>
<td>MiniRBT-H256</td>
<td>6</td>
<td>256</td>
<td>1024</td>
<td>8</td>
<td>10.4M (10.2%)</td>
<td>4.8M (5.6%)</td>
<td>6.8x</td>
</tr>
<tr>
<td>MiniRBT-H288</td>
<td>6</td>
<td>288</td>
<td>1152</td>
<td>8</td>
<td>12.3M (12.0%)</td>
<td>6.1M (7.1%)</td>
<td>5.7x</td>
</tr>
</tbody>
</table>

Table 3: Comparison of MiniRBT with baseline models on reading comprehension task and text classification task.

<table>
<thead>
<tr>
<th>Model</th>
<th>CMRC 2018 (F1/EM)</th>
<th>DRCD (F1/EM)</th>
<th>OCNLI (Acc)</th>
<th>LCQMC (Acc)</th>
<th>BQ Corpus (Acc)</th>
<th>TNEWS (Acc)</th>
<th>ChnSentiCorp (Acc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td>87.30/68.00</td>
<td>94.40/89.40</td>
<td>76.58</td>
<td>89.07</td>
<td>85.76</td>
<td>57.66</td>
<td>94.89</td>
</tr>
<tr>
<td>RBT6 (KD)</td>
<td>84.40/64.30</td>
<td>91.27/84.93</td>
<td>72.83</td>
<td>88.52</td>
<td>84.54</td>
<td>55.52</td>
<td>93.42</td>
</tr>
<tr>
<td>RBT3</td>
<td>80.30/57.73</td>
<td>85.87/77.63</td>
<td>69.80</td>
<td>87.30</td>
<td>84.47</td>
<td>55.39</td>
<td>93.86</td>
</tr>
<tr>
<td>RBT4-H312</td>
<td>77.90/54.93</td>
<td>84.13/75.07</td>
<td>68.50</td>
<td>85.49</td>
<td>83.42</td>
<td>54.15</td>
<td>93.31</td>
</tr>
<tr>
<td>MiniRBT-H256</td>
<td>78.47/56.27</td>
<td>86.83/78.57</td>
<td>68.73</td>
<td>86.81</td>
<td>83.68</td>
<td>54.45</td>
<td>92.97</td>
</tr>
<tr>
<td>MiniRBT-H288</td>
<td>80.53/58.83</td>
<td>87.10/78.73</td>
<td>68.32</td>
<td>86.38</td>
<td>83.77</td>
<td>54.62</td>
<td>92.83</td>
</tr>
</tbody>
</table>

Table 4: Comparing the results of two-stage distillation and one-stage distillation. The model is MiniRBT-H256 pre-trained with 30K steps, which is different from the published 100K step pre-trained model.

<table>
<thead>
<tr>
<th>Model</th>
<th>CMRC 2018</th>
<th>LCQMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>two-stage</td>
<td>77.97/54.60</td>
<td>86.58</td>
</tr>
<tr>
<td>one-stage</td>
<td>77.57/54.27</td>
<td>86.39</td>
</tr>
</tbody>
</table>

4.3 Results

Table 3 presents the results of MiniRBT’s performance on reading comprehension and text classification tasks. With only 10% of the parameters used by Chinese RoBERTa, MiniRBT achieves over 92% of its performance, a substantial improvement over RBT3 which has 3-4 times more parameters in reading comprehension tasks. In text classification tasks, MiniRBT achieves 98% of RoBERTa’s performance, with an average relative performance of 95.3% compared to Chinese RoBERTa. Table 3 further indicates that, with the same number of parameters (excluding the embedding layer), MiniRBT outperforms RBT4-H312, demonstrating that a narrow and deep model structure yields superior performance when compared to a wide and shallow model structure.

Table 4 reveals that two-stage pre-training distillation outperforms one-stage pre-training distillation in both reading comprehension and text classification tasks. These results suggest that the two-stage distillation approach effectively reduces the gap in size between teacher and student models, thus allowing students to maintain excellent performance even with small sizes.

5 Conclusion

In this study, we introduce MiniRBT, a small Chinese pre-trained model pre-trained with dynamic WWM, two-stage distillation, and a narrow and deep model structure. With only 10% of the parameters of RoBERTa, MiniRBT achieves an average relative performance of 94% and a speedup of 6.8x. Our findings demonstrate that MiniRBT has notable performance advantages over other models with the same number of parameters and can even surpass larger models with 3-4 times more parameters. In the future, we expect to combine pruning and quantization to propose more lightweight models.
References


Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert.


