MiniRBT：两阶段蒸馏小型中文预训练模型

姚鑫[†], Ziqing Yang[†], Yiming Cui[‡†], Shijin Wang[†]
[†]认知智能国家重点实验室，科大讯飞研究院，北京，中国
[‡]SCIR研究中心，哈尔滨工业大学，哈尔滨，中国
[†]{xinyao10,zqyang5,ymcui,sjwang3}@科大讯飞.com
[‡]ymcui@ir.hit.edu.cn

arXiv:2304.00717v1 [cs.CL] 3 Apr 2023

摘要

在自然语言处理中，预训练语言模型已成为重要的基础设施。然而，这些模型往往存在体积庞大、推理耗时长、难以部署等问题。此外，主流预训练模型大多专注于英语，针对小型中文预训练模型的研究仍然不足。本文介绍 MiniRBT，这是一种小型中文预训练模型，旨在推动中文自然语言处理研究。MiniRBT采用窄而深的学生模型，并在预训练期间结合全词掩码和两阶段蒸馏，使其能够胜任大多数下游任务。我们在机器阅读理解和文本分类任务上的实验表明，MiniRBT在性能上达到RoBERTa的94%，同时实现了6.8倍加速比，证明了其有效性和高效性。

1 引言

近年来，基于 Transformer（Vaswani等人，2017年）的预训练语言模型已成为自然语言处理的一种范式，其性能也随着模型规模的增长而不断提升。这些模型已主导自然语言处理和计算机视觉领域主要人工智能模型的榜单，包括 BERT（Devlin等人，2019年）、XL-Net（Yang 等人，2019年）、RoBERTa（Liu等人，2019年）、SpanBERT（Joshi等人，2020年）、ELECTRA（Clark等人，2020年）。尽管取得了显著进展，预训练模型在实际应用中仍面临诸多挑战，例如训练成本高、延迟大。此外，预训练模型的研究主要聚焦于英语，而针对中文的小型预训练模型则仍然缺乏。

因此，为进一步推动中文自然语言处理的发展，我们提出了一种实用性强的小型中文预训练模型。我们的方法采用动态全词掩码技术（Cui 等人，2021年）生成训练样本，以便全面建模粗粒度语义。我们为学生模型采用了窄而深的模型结构。在预训练过程中，我们采用包含教师助理模型的两阶段蒸馏方法。首先将教师模型的知识蒸馏到教师助理模型，再从教师助理模型蒸馏到学生模型。由此得到的模型更轻量、速度更快，可针对多个下游任务进行微调，并展现出优异的性能。

我们的研究主要致力于开发可应用于实际任务的小型中文预训练模型。本文的主要贡献可总结如下：

1. 我们通过实验得出结论：对于规模相近的网络，较窄且较深的结构比宽而浅的结构更有效。

2. 基于上述发现，我们提出了 MiniRBT[1]，，一种中文小型预训练模型，其规模仅为中文 RoBERTa 的 10%，同时平均性能达到 RoBERTa 的 94%。此外，其加速比达到 6 至 7 倍。

2 背景

2.1 中文 RoBERTa-wwm

自 BERT 出现以来，预训练语言模型取得了显著而迅速的发展。然而，传统的中文预训练模型（如最初的中文 BERT）通常采用一种分词方法，以单个字符为粒度对中文序列进行切分，这意味着中文词语

[1]可在此获取https://github.com/iflytek/MiniRBT

掩码策略	示例
原文	使用语言模型来预测下一个词的概率。
分词	使用 语言 模型 来 预测 下 一个 词 的 概率 。
原始掩码	使用语言 [MASK] 型来 [MASK] 测下一个词的概 [MASK] ##lity。
WWM	使 用 语 言 [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的 [MASK] [MASK] [MASK]。

表1：不同掩码策略示例。

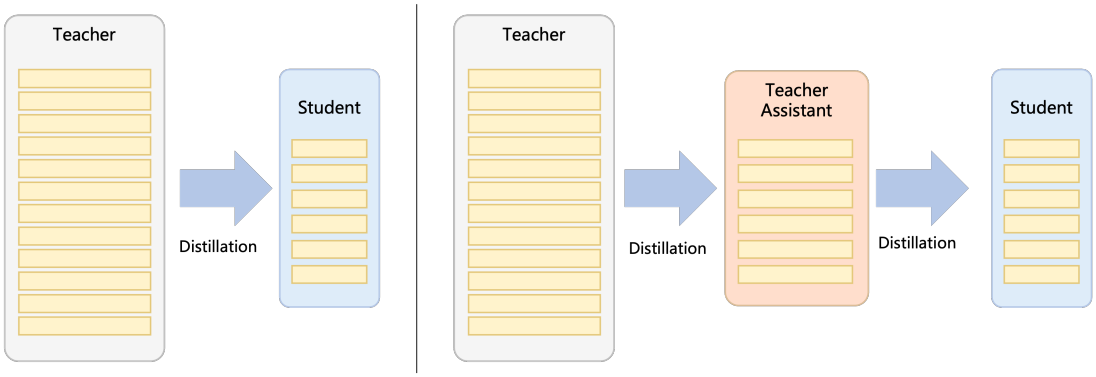


图1：单阶段蒸馏（左）与两阶段蒸馏（右）过程的对比。

会通过基于 WordPiece 的分词方法切分为字符。在生成训练样本时，这些字符会被逐个随机掩码，而不考虑中文分词。为解决这一问题，中文 RoBERTa-wwm（Cui 等人，2021年）采用了一种专为中文设计的全词掩码（WWM）方法。当一个中文词语中的部分内容被掩码时，该词语的其他部分也会同时被掩码。需要注意的是，WWM 方法只影响预训练阶段掩码词元的选择。中文 RoBERTa-wwm 使用中文维基百科（包括简体中文和繁体中文）进行训练。在我们的工作中，我们使用中文 RoBERTa-wwm 作为教师模型。

2.2 知识蒸馏

近年来，越来越多的研究提出了模型压缩方法，以减少模型参数数量并提高模型推理速度，例如量化（Shen 等人，2020年；Fan 等人，2020年）、剪 Xia 等人，2022年；Lagunas 等人，2021年）、知Jiao 等人，2020年；Sun 等人，2020年；Sanh 等人，2019年；Hou 等人，2020年）。知识蒸馏（KD）通过训练学生模型模仿教师模型的行为，将大型教师模型中蕴含的知识迁移到小型学生模型中。例如，TinyBERT（Jiao 等人，2020年）首先在通用领域的大规模语料库上进行通用蒸馏，得到一个通用蒸馏小模型，然后在微调阶段利用下游数据进行特定任务的蒸馏。

DynaBERT（Hou 等人，2020年）通过将全尺寸模型的知识蒸馏到小型子网络中，训练出宽度和深度均可自适应的 BERT。知识蒸馏也应用于剪枝以提升性能，例如块剪枝（Lagunas 等人，2021年）和 CoFi（Xia 等人，2022年）。

3 方法

本节介绍 MiniRBT 所采用的相关方法。

3.1 全词掩码

原始的基于 WordPiece 的分词方法会将中文输入序列拆分为彼此独立的字符，并随机掩码这些字符。该方法简化了预测过程，因为模型只需记住词语中特定的字符顺序，就能预测被掩码的字符，而无需考虑语义上的上下文关系。为提升模型性能，我们在预训练过程中采用动态全词掩码（WWM）方法作为训练样本生成策略。首先使用传统的中文分词（CWS）方法将输入切分为中文词语，随后按照全词掩码（WWM）进行掩码，如表1所示。我们使用 LTP（Che 等人，2010年）作为首选工具来提取词语

分割信息。

3.2 两阶段蒸馏

传统的知识蒸馏方法直接将知识从教师模型传递给学生模型。然而，当教师模型和学生模型的结构存在显著差异时，这种方法可能导致性能差距。为解决这一问题，我们提出在预训练阶段采用两阶段蒸馏方法，该方法基于教师助理知识蒸馏的理念（Mirzadeh 等人，2020年）。如图1所示，该方法先将知识从教师模型（RoBERTa）蒸馏至教师助理（RBT6），再从教师助理蒸馏至学生模型（MiniRBT）。教师助理这一中间环节有助于缩小教师模型与学生模型之间的规模差距，进而提升学生模型在下游任务中的表现。

为使用隐藏层蒸馏和预测层蒸馏进行知识蒸馏（KD），我们采用了 TextBrewer（Yang 等人，2020年），这是一款基于 PyTorch、专为自然语言处理设计的模型蒸馏工具包。我们从隐藏层的输出中蒸馏知识。目标是

$$\mathcal{L}_{\text{layer}} = \sum \text{MSE}(H_s^{i'} W_h, H_t^i) \quad (1)$$

其中，$H_s^{i'} \in \mathbb{R}^{l \times d'}$ 和 $H_t^i \in \mathbb{R}^{l \times d}$分别表示第 i' 个学生模型隐藏层和第 i 个教师模型隐藏层的隐藏表示。$W_h \in \mathbb{R}^{d' \times d}$ 是一个线性变换，用于匹配学生网络的隐藏状态和教师网络的隐藏状态。除了模仿教师的隐藏层行为外，我们还使用教师模型的软目标概率，通过交叉熵损失训练学生模型

$$\mathcal{L}_{\text{pred}} = -p(z^T) \cdot \log p(z^S) \quad (2)$$

其中，z^S 和 z^T 分别是学生模型和教师模型预测的 logits 向量，而 p =softmax(z/t)是经过温度 t 缩放的概率，其中 logits 为 z。

最后，我们将隐藏层蒸馏与预测层蒸馏相结合：

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{layer}} + \mathcal{L}_{\text{pred}} \quad (3)$$

3.3 更窄更深的学生模型

通过初步实验，我们发现，在参数量相同的情况下，窄而深的模型结构优于宽而浅的结构。因此，我们为 MiniRBT 采用了窄而深的设计。模型结构的详细信息见表2。MiniRBT 包含两个模型分支：MiniRBT-H256 和 MiniRBT-H288。这些模型的隐藏层维度分别为 256 和 288，并包含 6 个 Transformer 层，采用两阶段蒸馏方法进行预训练。

4 实验

4.1 下游任务

机器阅读理解 机器阅读理解（MRC）是一项文档级建模任务，要求模型根据给定段落回答问题。我们在两个中文阅读理解数据集上评估了模型：CMRC 2018（Cui 等人，2019年）和 DRCD（Shao 等人，2018年）。它们与 SQuAD（Rajpurkar 等人，2018）的形式相似。评估指标为 F1 和精确匹配（EM）。

文本分类 对于单句分类，我们使用 TNEWS 和 ChnSentiCorp（Tan 和 Zhang，2008）。ChnSentiCorp 数据集涉及情感分类，文本需分为正面或负面；TNEWS 数据集则涉及将短文本划分为不同的新闻类别。对于句对分类，我们选取了三个数据集：OCNLI、LCQMC（Liu等人，2018年）和 BQ语料库（Chen 等人，2018年）。OCNLI 和 TNEWS 均被纳入中文语言理解测评基准（CLUE）的子任务（Xu 等人，2020年）。这些任务的评估指标为准确率。

4.2 训练设置

在预训练阶段，训练批次大小设为 4096，峰值学习率设为 4e-4，温度设为 8，训练步数为 100K。

我们分别在下游任务上对模型进行 2、3、5 和 10 个训练轮数的微调，学习率从 $\{5e-5, 1e-4\}$ 中选择。为降低随机性对实验结果的影响，我们使用不同的随机种子对每项任务至少运行三次，并报告平均性能

模型	层数	隐藏层维度	前馈网络维度	注意力头数	模型大小	模型大小（不含嵌入层）	加速比
RoBERTa-wwm RBT6（知识蒸馏）	12	768	3072	12	102.3M (100%)	85.7M (100%)	1x
	6	768	3072	12	59.8M (58.4%)	43.1M (50.3%)	1.7x
RBT3	3	768	3072	12	38.5M (37.6%)	21.9M (25.6%)	2.8x
RBT4-H312	4	312	1200	12	11.4M (11.1%)	4.7M (5.5%)	6.8x
MiniRBT-H256	6	256	1024	8	10.4M (10.2%)	4.8M (5.6%)	6.8x
MiniRBT-H288	6	288	1152	8	12.3M (12.0%)	6.1M（7.1%）	5.7x

表 2：模型结构比较。RBT 代表 RoBERTa，RBT3 以 RoBERTa 的前三层初始化，并进行了 100 万步预训练。层数不包括嵌入层和预测层。

Task	CMRC 2018 （F1/精确匹配）	DRCD （F1/精确匹配）	OCNLI （准确率）	LCQMC （准确率）	BQ语料库 （准确率）	TNEWS （准确率）	ChnSentiCorp （准确率）
RoBERTa RBT6（知识蒸馏）	87.30/68.00	94.40/89.40	76.58	89.07	85.76	57.66	94.89
	84.40/64.30	91.27/84.93	72.83	88.52	84.54	55.52	93.42
RBT3	80.30/57.73	85.87/77.63	69.80	87.30	84.47	55.39	93.86
RBT4-H312	77.90/54.93	84.13/75.07	68.50	85.49	83.42	54.15	93.31
MiniRBT-H256	78.47/56.27	86.83/78.57	68.73	86.81	83.68	54.45	92.97
MiniRBT-H288	80.53/58.83	87.10/78.73	68.32	86.38	83.77	54.62	92.83

表3：对比　MiniRBT 与基线模型在阅读理解任务和文本分类任务上的表现　　sk.

模型	CMRC 2018	LCQMC
两阶段	**77.97/54.60**	**86.58**
单阶段	77.57/54.27	86.39

表4：比较两阶段蒸馏和单阶段蒸馏的结果。该模型为经过3万步预训练的 MiniRBT-H256，与已发布的经过10万步预训练的模型不同。

开发集上的分数。值得注意的是，对于较小的模型，增加迭代次数和学习率往往能提升其在下游任务上的性能。

模型加速比是相对于单块 NVIDIA M40 GPU 上的中文 RoBERTa 进行评估的。所有任务的输入长度均设为 512，批量大小为 128。

4.3 结果

表3 展示了 MiniRBT 在阅读理解和文本分类任务中的表现。MiniRBT 仅使用中文 RoBERTa 10% 的参数量，便达到了其 92% 以上的性能；在阅读理解任务中，相较于参数量多出 3 至 4 倍的 RBT3，取得了显著提升。在文本分类任务中，MiniRBT 达到 RoBERTa 98% 的性能；与中文 RoBERTa 相比，平均相对性能为 95.3%。表 3 还进一步表明，在参数量相同的情况下（不包括嵌入层），MiniRBT 的表现优于 RBT4-H312，这表明，与宽而浅的模型结构相比，窄而深的模型结构能够带来更优的性能。

表4 显示，两阶段预训练蒸馏在阅读理解和文本分类任务上的表现均优于单阶段预训练蒸馏。这些结果表明，两阶段蒸馏方法有效缩小了教师模型与学生模型之间的规模差距，使学生模型即使规模较小，也能保持出色的性能。

5 结论

本研究介绍了 MiniRBT，这是一种小型中文预训练模型，采用动态 WWM、两阶段蒸馏以及窄而深的模型结构进行预训练。MiniRBT 的参数量仅为 RoBERTa 的 10%，平均相对性能达到 94%，加速比为 6.8 倍。我们的研究结果表明，与参数量相同的其他模型相比，MiniRBT 具有显著的性能优势，甚至能够超越参数量多出 3–4 倍的大型模型。未来，我们计划结合剪枝和量化，提出更加轻量化的模型。

参考文献

Wanxiang Che、Zhenghua Li 和 Ting Liu。2010年。LTP：中文语言技术平台。发表于*Coling 2010*：演示，第13–16 页。

Jing Chen、Qingcai Chen、Xin Liu、Haijun Yang、Daohe Lu 和 Buzhou Tang。2018年。BQ 语料库：用于句子语义等价识别的大规模领域专用中文语料库。载于*2018*年自然语言处理实证方法会议论文集，第 4946–4951 页，比利时布鲁塞尔。计算语言学协会。

Kevin Clark、Minh-Thang Luong、Quoc V Le 和 Christopher D Manning。2020年。ELECTRA：将文本编码器预训练为判别器而非生成器。*arXiv* 预印本 *arXiv:2003.10555*。

Yiming Cui、Wanxiang Che、Ting Liu、Bing Qin 和 Ziqing Yang。2021年。采用全词掩码对 ChineseBERT 进行预训练。

Yiming Cui、Ting Liu、Wanxiang Che、Li Xiao、Zhipeng Chen、Wentao Ma、Shijin Wang 和 Guop-ing Hu。2019年。用于中文机器阅读理解的跨度抽取数据集。载于*2019*年自然语言处理实证方法会议暨第九届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第 5883–5889 页，中国香港。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：面向语言理解的深度双向变换器预训练。载于：*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第 *1* 卷（长文与短文），第 4171–4186 页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Angela Fan、Pierre Stock、Benjamin Graham、Edouard Grave、Rémi Gribonval、Herve Jegou 和 Armand Joulin。2020年。使用量化噪声进行极端模型压缩。*arXiv* 预印本 *arXiv:2004.07320*。

Lu Hou、Zhiqi Huang、Lifeng Shang、Xin Jiang、Xiao Chen 和 Qun Liu。2020年。DynaBERT：具有自适应宽度与深度的动态 BERT。神经信息处理系统进展，33：9782–9793。

Xiaoqi Jiao、Yichun Yin、Lifeng Shang、Xin Jia ng、Xiao Chen、Linlin Li、Fang Wang 和 Qun Liu。2020年。TinyBERT：蒸馏 BERT，以实现自然语言理解。载于计算语言学协会论文集：*EMNLP 2020*，第4163–4174 页，线上。计算语言学协会。

Mandar Joshi、Danqi Chen、Yinhan Liu、Daniel S. Weld、Luke Zettlemoyer 和 Omer Levy。2020年。SpanBERT：通过表示改进预训练并预测文本跨度。计算语言学协会汇刊，8:64–77。

François Lagunas、Ella Charlaix、Victor Sanh 和 Alexander Rush。2021年。用于加速 Transformer 的块剪枝。载于*2021*年自然语言处理实证方法会议论文集，第 10619–10629 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Xin Liu、Qingcai Chen、Chong Deng、Huajun Zeng、Jing Chen、Dongfang Li 和 Buzhou Tang。2018年。LCQMC：大规模中文问题匹配语料库。载于第*27*届计算语言学国际会议论文集，第 1952–1962 页。

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。

Seyed Iman Mirzadeh、Mehrdad Farajtabar、Ang Li、Nir Levine、Akihiro Matsukawa 和 Hassan Ghasemzadeh。2020年。通过教师助理改进知识蒸馏。载于*AAAI*人工智能会议论文集，第34卷，第5191–5198页。

Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。知其所不知：SQuAD 中无法回答的问题。载于第*56*届计算语言学协会年会论文集（第*2*卷：短篇论文），第784–789页，澳大利亚墨尔本。计算语言学协会。

Victor Sanh、Lysandre Debut、Julien Chaumond 和 ThomasWolf。2019年。Distilbert：BERT 的蒸馏版本，更小、更快、更便宜、更轻量。

Chih ChiehShao、Trois Liu、YutingLai、Yiying Tseng 和 Sam Tsai。2018年。Drcd：一个中文机器阅读理解数据集。*arXiv* 预印本 *arXiv:1806.00920*。

ShengShen、Zhen Dong、Jiayu Ye、Linjian Ma、Zhewei Yao、Amir Gholami、Michael W Mahoney 和 Kurt Keutzer。2020年。Q-BERT：基于Hessian矩阵的 BERT 超低精度量化。载于*AAAI*人工智能会议论文集，第34卷，第8815–8821页。

Zhiqing Sun、Hongkun Yu、Xiaodan Song、Renjie Liu、Yiming Yang 和 Denny Zhou。2020年。MobileBERT：适用于资源受限设备的紧凑型任务无关 BERT。载于计算语言学协会第*58*届年会论文集，第2158–2170页，线上。计算语言学协会。

Songbo Tan 和 Jin Zhang。2008。中文文档情感分析的实证研究。《专家系统与应用》，34(4):2622–2629。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。Attention Is All You Need。神经信息处理系统进展，30。

MengzhouXia、Zexuan Zhong 和 Danqi Chen。2022年。结构化剪枝能够学习出紧凑且准确的模型。载于计算语言学协会（*ACL*）。

Liang Xu、Hai Hu、Xuanwei Zhang、Lu Li、Chenjie Cao、Yudong Li、Yechen Xu、Kai Sun、Dian Yu、Cong Yu 等。2020年。中文语言理解测评基准：中文语言理解评测基准。*arXiv* 预印本 *arXiv:2004.05986*。

Zhilin Yang、Zihang Dai、Yiming Yang、Jaime Carbonell、Russ RSalakhutdinov 和 Quoc VLe。2019年。XLNet：用于语言理解的广义自回归预训练。神经信息处理系统进展，32。

Ziqing Yang、Yiming Cui、Zhipeng Chen、Wanxiang Che、Ting Liu、Shijin Wang 和 Guoping Hu。2020年。TextBrewer：用于自然语言处理的开源知识蒸馏工具包。载于第*58*届计算语言学协会年会论文集：系统演示，第9–16页，线上。计算语言学协会。