文章编号：2095-2163(2023)01-0129-07 中图分类号：TP391 文献标志码：A

SmBERT(SmallerBert)：一种更小更快的文本分类模型

王 淼[1]，丁德锐[2]
(1 上海理工大学 光电信息与计算机工程学院，上海 200093；2 上海理工大学 理学院，上海 200093)

摘 要：文本分类是自然语言处理的一个重要领域，随着深度网络的发展，大规模预训练模型日渐成为文本分类任务的主流模型，但大模型的推理速度慢、尺寸大难以在计算资源有限的设备应用，而且大模型多存在参数冗余问题。为了在不损失过多性能的情况下尽可能地预训练大模型进行模型压缩，本文提出了一种更小更快的 BERT 类模型 SmBERT。该模型由原生 BERT 首先经过二次自蒸馏，纵向实现二倍压缩率；其次，经过多学习目标的知识蒸馏，多维度迁移大模型的语言知识，从而丰富目标模型的语言理解能力；最后，使用面向跨语言任务的剪枝，从隐层和注意力头方向实现模型的宽度剪枝，最终得到 SmBERT。通过测试，在 QQP、QNLI、SST-2、MRPC 和 RTE 数据集上，只有 BERT 的 35%参数量的 SmBERT 表现了其 94%的综合性能，并在小数据集 RTE 上超越了 BERT 模型，推理速度提升了 6.1 倍。

关键词：文本分类；模型压缩；SmBERT；BERT；知识蒸馏

SmBERT(SmallerBert): a smaller and faster class model of text classification

WANG Miao[1], DING Derui[2]
(1 School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2 School of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 Text classification is an important field of natural language processing, in recent years, with the development of deep networks, large-scale pre-trained models have increasingly become the mainstream model of text classification tasks, but the inference speed of large models is large, and the size is difficult to apply in devices with limited computing resources, and large models have parameter redundancy problems. In order to perform model compression on pre-trained models as much as possible in the case of excessive performance loss, this paper proposes a smaller and faster BERT class model SmBERT, which is double the compression ratio by the native BERT first through secondary self-distillation longitudinally, followed by the knowledge distillation of multiple learning targets, and then the language knowledge of the large model is enriched by multi-dimensional migration of the language knowledge of the target model, and finally the width of the model is pruned from the hidden layer and attention head direction using cross-language task pruning to achieve the width of the model and finally obtain SmBERT. Through testing, SmBERT, which has only 35% of the parameter amount of BERT on the QQP, QNLI, SST-2, MRPC and RTE datasets, performed 94% of its comprehensive performance and surpassed the original model on the small dataset RTE, and the inference speed was increased by 6.1 times.

【Key words】 text classification; model compression; BERT; knowledge distillation

0 引 言

越来越多的信息以文本的形式存储，因此文本挖掘具有很高的商业价值。文本分类是文本挖掘的重要任务，是从文本信息中提取知识的技术，广泛应用于医学、社会科学、商业等领域。而手动设计特征的手工文本分类方法难以处理海量的文本信息，因此基于深度学习的方法越来越多的被应用文本分类任务中[1]。

近年来，基于自注意力机制的预训练语言模型 BERT，因其良好的准确性和预训练-微调机制的多任务通用性，逐渐成为自然语言理解和生成领域的主流模型[2]。例如：采取动态掩码和更长序列训练的 Roberta，结合双流注意力机制和置换语言模型(Permutation Language Model)的自回归语言模型 XLNet，将孪生网络引入 BERT 网络生成具有较强通

作者简介：王 淼(1996-)，男，硕士研究生，主要研究方向：文本分类、自然语言处理；丁德锐(1981-)，男，博士，教授，博士生导师，主要研究方向：分布式优化控制与滤波、图像处理与智能计算。
通讯作者：丁德锐 Email：deruiding2010@usst.edu.cn

收稿日期：2022-04-05

用性的句向量的 SentenceBERT,在少样本(few-shot)、单样本(one-shot)和零样本(zero-shot)任务均取得了令人震惊的效果的 GPT-3。

上述模型在文本分类任务中都取得了良好的效果,但由于其参数量巨大,其中 GPT-3 的参数量甚至达到了 1 700 亿之多,且自注意力机制的计算时间复杂度与输入序列的平方成正比,因此模型训练成本高昂,内存需求和功耗巨大,预测速度慢,难以部署在移动端[3]。现实世界的移动端或资源受限的设备,往往需要小尺寸、高精度、快速推理、低电耗的模型,严重限制了 BERT 类模型在实际生活中的落地和发展。一些相关工作证明了 BERT 类模型网络权重存在参数冗余的特征,尤其是对于文本分类的自然语言理解式任务,参数冗余问题就更加凸显[4-5]。这给对 BERT 类模型进行模型压缩提供了理论支撑。

为了压缩模型,本文提出了一种基于自蒸馏、多目标知识蒸馏和剪枝的小型文本分类模型 SmBERT,通过二次自蒸馏逐步将原生 12 层 BERT 模型压缩至 3 层;对得到的 3 层 BERT 进行多目标知识蒸馏,进而通过多尺度融合学习教师模型的知识;最后,通过剪枝对模型进行纵向压缩以得到 SmBERT 网络。

1 相关工作

随着语言模型的改进和大规模预训练模型的兴起,文本分类在最近几年取得了飞速发展。但随着模型的规模不断扩大,模型的训练成本和运行硬件要求都日益提高,难以在计算能力有限的设备上运行。同时大模型往往存在参数冗余的问题,面向小数据集任务时,冗余问题则越发凸显,近年来一些相关工作通过随即丢弃、剪枝和稀疏矩阵等方法对模型进行压缩,探究在不损失过多精度的条件下尽可能地对模型进行压缩[6]。

1.1 自注意力机制

注意力机制起源于人类在较大的感受野中利用有限注意力提取重要信息的视觉处理机制,可以快速提取全局信息的重要特征。

注意力机制最初应用在计算机视觉的相关任务,并取得了较好的结果。而在自然语言处理领域中,由于语言天生从左向右单向排列的特性,应用广泛的传统的语言模型如 RNN,LSTM,GRU 等均是单向自回归语言模型,不具备并行处理能力:

(1)只能逐字输入,效率极低;

(2)无法同时利用上下文信息,难以捕捉长距离文本的信息;

(3)存在“长期依赖”和梯度消失问题。因此严重限制了自然语言领域的发展,也限制着注意力机制在自然语言领域中的应用。

2017 年,Vaswani[7] 等人提出了改进的自注意力机制(Self-Attention),通过在输入数据内部建立特征连接,可以并行的处理一整段文字而不是逐词按时序处理,增强了长距离上下文信息的捕捉能力和对数据内部相关性特征的提取能力,减弱了对额外信息的需求。

1.2 模型剪枝

1989 年,LeCun Y[8] 提出“最优脑损伤”(Optimal Brain Damage)的剪枝压缩方法,可以通过类似生物突触剪枝的方法来缓解网络“过度能力(过拟合)”的问题。

最初的模型剪枝多针对权重连接和神经元单元的非结构剪枝方法,在训练过程中剪除不重要的连接和神经元,衡量重要性的标准多是根据参数绝对值的大小。

随着深度学习的发展,神经网络层数不断加深,引起了梯度消失和梯度爆炸问题,限制着深度网络的发展,而针对模型层级的结构性剪枝通过对层数的修剪,在不损失过多性能的情况下降低了深度模型的复杂度。

1.3 知识蒸馏

受启发于人类学习的方式,2015 年 Hinton[9] 提出了知识蒸馏(Knowledge Distillation),核心思想是让学生小模型模仿教师模型,从而学习大模型的通用语言知识。

2 模型框架

本文发现通过基于目标的知识蒸馏来获得 3 层的 BERT 子网络,其模型效果甚至不如原生 BERT 网络的前 3 层子网络,说明了单一蒸馏方法的局限性。简单初始化的学生网络难以有效地学习教师网络的知识,而单一的知识蒸馏目标难以充分的迁移教师网络的知识。如果采用多目标的知识蒸馏手段,同时利用具有较好泛化性的教师模型来完成学生网络的参数初始化,上述问题可以得到一定程度的解决。基于此,本文提出了基于自蒸馏、多目标知识蒸馏和剪枝的 SmBERT 网络,该网络由原生 12 层的 BERT-base 通过两次自蒸馏压缩层级,得到 3 层的子网络,随后应用多目标知识蒸馏再次将原生

BERT 模型的知识迁移到学生模型中，最后通过面向任务的剪枝进行注意力头的剪枝，最终得到 SmBERT 网络，其网络结构如图 1 所示。

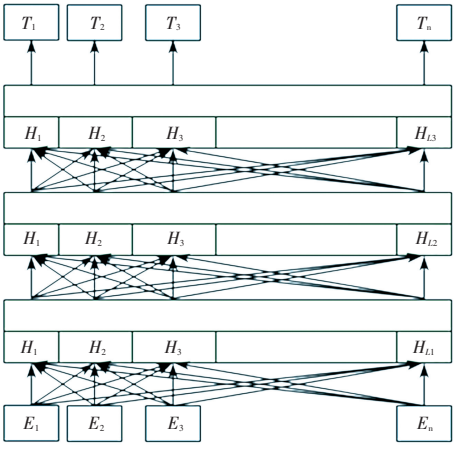

图 1　SmBERT 网络结构

Fig. 1　SmBERT network structure

通过自蒸馏的方式，在不引入教师模型的前提下，通过不断地用子网络的模块替换母网络的相应模块，使子网络逐渐拟合并代替母网络；相比于传统的单目标知识蒸馏，本文的多目标学习学生模型从更多维度的学习教师模型的知识；相比于传统的基于单一任务剪枝，本文的多语言任务剪枝能使模型学习更通用的语言知识。

2.1　BERT 网络

2018 年，Devlin 提出基于自注意力机制的 BERT 模型，在 GLUE 数据集和 SQuAD 数据集的 11 个自然语言理解子任务中取得了 SOTA(State-of-the-Art)的结果，证明了基于自注意力的双向语言模型的有效性，同时参数量巨大的 BERT 模型提取了目标文本的通用词向量表示，从而使其网络后外接面向当前任务的分类层并微调就可以广泛应用于自然语言理解的不同子任务中，这使得 BERT 类预训练-微调的模型训练范式日趋流行。

BERT 通过多头自注意力机制实现对输入文本所有位置字词的相关性计算，使得每个词的向量表征同时抽取到上下文信息。BERT 的输入处理如公式(1)所示，通过 Embedding 词嵌入处理，得到分布式词向量矩阵 $\boldsymbol{X} \in R^{d_{model} \times d_{length}}$。

$$\boldsymbol{X} = Embedding(input) \tag{1}$$

其中，$input$ 为文本。

$\boldsymbol{X}$ 与 3 个不同权值的嵌入矩阵 $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V$ 相乘，得到输入序列对应的 $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 投影矩阵，公式(2)和公式(3)：

$$(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{X} \times (\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V) \tag{2}$$

$$d_k = d_{model} \div head \tag{3}$$

其中，d_{model} 为词向量矩阵的行数，$head$ 为注意力头的数量。

为了计算输入文本中每个位置字词的自注意力分数，矩阵左乘 $\boldsymbol{K}$ 的转置，以建立输入序列的自注意力矩阵，除以分布的标准差 $\sqrt{d_k}$ 以维持数据依然保持近似的正态分布，从而维持梯度稳定，并应用 softmax 归一化将分数转化为概率。最后与 $\boldsymbol{V}$ 矩阵相乘，得到融合输入序列自相关性信息的新的词向量矩阵，为公式(4)

$$\boldsymbol{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{4}$$

将每个自注意力头得到的矩阵拼接并投影，得到最终输出的词向量矩阵，为公式(5)

$$\boldsymbol{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_1, \cdots, head_2\ W^o) \tag{5}$$

其中，$head_i = Attention(QW_i^Q,\ KW_i^K, VW_i^V)$；Concat 表示将这 i 个注意力头的输出矩阵拼接；$\boldsymbol{W}^o$ 为随机初始化的矩阵。

BERT 通过自注意力机制获得文本内全局特征信息，在多任务中都表现了超越传统模型的效果，启发了一系列相关改进和探究工作，但 BERT 模型的巨大参数量、深度以及自注意力机制序列长度二次方的时间复杂度都严重限制了 BERT 在小设备的应用。

针对上述问题，可通过稀疏化自注意力矩阵来降低模型的时间复杂度，将自注意力机制的时间复杂度改进为近似于序列长度的线性关系，如 Longformer，Big Bird 等通过稀疏化自注意力矩阵，将自注意力机制的时间复杂度改进为近似于序列长度的线性关系，但这种稀疏矩阵往往需要单独建立底层运算的稀疏运算加速库。与此同时，诸如剪枝，知识蒸馏，权重共享及量化等模型压缩方法越来越多的被用于改善 BERT 类网络的速度，从模型的深度和宽度两个维度上对模型进行压缩，但权重共享只能缩小模型尺寸而无法实现运行的加速，量化策略牺牲模型的精度，因此本文采用多目标知识蒸馏，自蒸馏及剪枝同时从深度和宽度压缩原生 BERT 模型。

2.2　自蒸馏

受忒修斯之船悖论的启发，Canwen Xu 于 2019 年提出 BERT of Theseus 网络(以下简称 Theseus)，

将目标模型分为多个模块。通过自蒸馏的方法将 6 层的目标模型压缩至 3 层。首先利用目标模型的前 3 层来初始化子模型,将母模型每两层视为一个母模块 Prd_i, 每层子模型视为一个子模块 Scc_i, 训练过程如图 2 所示。

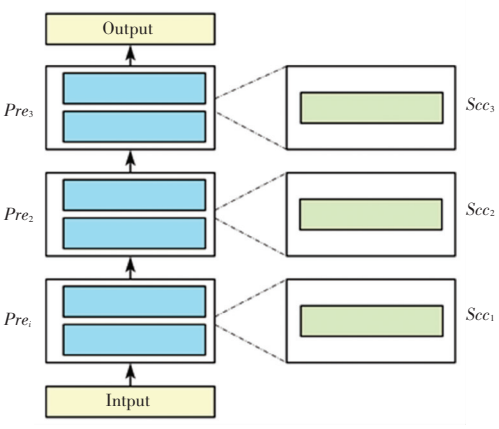

图 2 Theseus 网络训练过程

Fig. 2 Theseus training process

训练的方式有两种:第一种固定概率蒸馏,即在训练的时候让子模块 Scc_i 以既定的替换概率 P 替换相对应的母模块 Pre_i, 子模块在逐渐替换训练的过程中达到了母模块在整体模型中的效果,在训练一定步长后结束蒸馏,为公式(6)所示

$$Pre_{i+1} = P \times Scc_i + (1 - P) \times Pre_i \tag{6}$$

其中, P 表示人为设定的替换率; Scc_i 表示第 i 个子模块;Pre_i 表示第 i 个母模块。

第二种是调节概率蒸馏,即在蒸馏的过程中逐渐增加替换的概率,最终替换概率达到 1,为公式(7),即所有子模块替换对应的母模块并组成自蒸馏后的压缩子模型,从而将深度压缩到原来的一半。

$$P = \min(1, kt + b) \tag{7}$$

其中, k 与 b 均为人为设置的超参数,t 表示训练步长。

本文采用调节概率蒸馏法。

Theseus 网络采用基于具体任务的蒸馏方式,可以直观地观察蒸馏的有效性,但只是从深度这一单一维度上进行压缩,横向依旧存在参数冗余。因此,本文采用剪枝的手段进行宽度压缩。

2.3 基于多目标训练的知识蒸馏

随着深度学习的发展,复杂的大模型日益成为主流,需要大量的计算资源以便从高维冗余的语料库中学习通用的语义知识,但同时对部署设备的要求很高。为了利用大模型来提升小模型的效果,传统的知识蒸馏训练过程分为教师模型的训练和学生模型的学习两个阶段。

对于分类任务而言,教师模型的训练大致与一般网络的训练过程无异,只是一般神经网络通过 softmax 函数计算输出分类概率分布,其计算为公式(8)

$$\mathrm{softmax}(Z_i) = \frac{e^{Z_i}}{\sum_i e^{Z_j}} \tag{8}$$

其中, Z_i 是网络输出第 i 类的逻辑值。

但除了正标签外,负标签的分布也蕴含着语义信息。而复杂大模型输出的分布向量近似于正标签为 1,负标签接近于 0 的独热向量,难以从输出概率分布中传递知识,因此教师模型引入了温度 T 来产生软化的标签,以改良传统的 softmax 函数,模型新的输出分类概率分布为公式(9),期以提高学生模型的泛化能力。

$$p_i = \frac{e^{(Z_i/T)}}{\sum_i e^{(Z_j/T)}} \tag{9}$$

学生模型的训练目标函数是结合软标签和真实标签的交叉熵函数,在学习教师模型知识的同时引入真实分布知识,即学生模型既学习教师模型的预测结果,也学习样本的真实标签,为公式(10)

$$L_{ce} = a \times \sum_i t_i \times \log s_i + b \times \sum_i f_i \times \log t_i \tag{10}$$

其中, t_i 和 s_i 分别表示教师模型和学生模型对第 i 类的预测概率值;f_i 表示第 i 类的真实标签值; a 和 b 是人为设定的超参数,分别代表了软标签和真实标签对损失函数的贡献程度。

然而基于交叉熵损失的知识蒸馏方法只利用了网络最终的输出,没有多维度地蒸馏网络内部知识,提升效果一般。

为了多维度地学习教师模型的知识,本文借鉴了 DistilBERT 的蒸馏方式,采用了多目标的蒸馏方法,除了传统的 交叉熵损失函数 L_{ce} 外,本文也采用了其他损失函数作为学习目标,如掩码语言模型和余弦相似度。

L_{mlm} 是 Masked Language Model(掩码语言模型) BERT 自监督预训练任务的损失函数,用当前被掩盖词的前后文来预测该词,使得模型能够有效地学习双向编码信息。为了从无标签的大数据集中获取通用语言知识,本文采用掩码语言模型损失函数 L_{mlm} 作为损失函数,计算为公式(11)

$$L_{mlm} = -\sum_{\hat{x} \in m(x)} \log p(\hat{x}) \mid x_{\setminus m(x)} \tag{11}$$

其中, $m(x)$ 表示输入序列被掩盖的字词;

$x_{\backslash m(x)}$ 表示未被掩盖的字词；$P(\hat{x} \mid x_{\backslash m(x)})$ 表示根据输入序列中未被掩盖的字词推理出真实被掩盖字词的概率。

为了利用教师模型的隐层知识，本文还采用了教师和学生模型隐层的余弦相似度 $L_{\cos}$ 作为知识蒸馏的目标之一，其值越大则说明学生模型和教师模型的隐层语义空间越接近，蒸馏的效果越好，来调整教师和学生模型的隐层参数方向，为公式(12)

$$L_{\cos} = Weight_{teacher} \cdot Weight_{student} \tag{12}$$

其中，$Weight_{teacher}$ 和 $Weight_{student}$ 分别表示教师模型和学生模型的隐层参数。

综上，本文采用的多目标知识蒸馏的损失函数为公式(13)

$$Loss = \alpha \times L_{ce} + \beta \times L_{mlm} + \gamma \times L_{cos} \tag{13}$$

其中，α、β 和 γ 均是人为设置的超参数，分别代表了 L_{ce}、L_{mlm} 和 $L_{\cos}$ 对 $Loss$ 函数的贡献程度。

2.4　多维剪枝

BERT 类模型采用多头自注意力机制，由若干个维度较低的注意力头组成，形成多个子空间，以捕捉多层次的语义关系。而 Michel[10] 经过头剪枝消融实验证明了多头注意力机制存在冗余：在每层只丢弃一个注意力头的情况下，大部分时候模型的性能得到略微提升，说明多头注意力机制存在注意力头冗余的问题。

BERT 类模型还存在隐层维度不对齐的问题，例如 BERT-base 的隐层维度为 768，而 BERT-large 的隐层维度为 1 024，这就限制了低维度的小模型学习高维度模型的能力，Sun[11] 提出瓶颈(bottleneck)机制，即对齐大小模型的输入输出维度，而在隐层内部又将中间输出维度还原为输入维度，从中间隐层维度的方向进行模型压缩。

使用目标数据集对模型进行训练，模型中绝对值越小的参数往往对当前任务具有较少的贡献，因而其被剪枝的优先级则越高，为公式(14)

$$Top_v(Weight_i) = \begin{cases} Weight_i, Weight_i in\ top\ v\% \\ 0, \quad otherwish \end{cases} \tag{14}$$

其中，top 表示原模型参数，Top_v 表示绝对值大小排名前 $v\%$ 的参数。

本文使用哈工大讯飞实验室开发的 Text-Pruner 程序包对模型进行注意力头和隐层维度的横向压缩。

3　实　验

为评估本文提出 SmBERT 模型的性能，在 GLUE (The General Language Understanding Evaluation)的 QQP、MRPC、RTE、QNLI 和 SST2 这 5 个子任务上进行了测试，并与原生 BERT 及其一系列衍生模型进行了比较。计算机配置为：Intel(R) Xeon(R) CPU E5-1620 v3 @ 3.50 GHz，Nvidia GTX 1080 8 GB RAM。

在 GLUE 的 MNLI 数据集上对模型进行二次自蒸馏，两次蒸馏的参数配置见表 1。

表 1　两次自蒸馏的训练参数

Tab. 1　Training parameters for two self-distillations

参数	第一次蒸馏	第二次蒸馏
初始替换率 R	0.3	0.4
概率增长率 K	0.000 6	0.000 8
学习率	0.000 01	0.000 01
迭代周期	15	15

第一阶段将原生 BERT 蒸馏到 6 层，并对目标模型进行第二阶段的蒸馏，将层数降低为 3 层。鉴于第一阶段的母模型参数量较大，本文采用了较低的初始替换率 R 和较低的概率增长率 K，让子模型在蒸馏的过程中学习曲线更加平滑；与此对应的是第二阶段的蒸馏，本文采用了较大的替换率和增长率。

继续在 MNLI 数据集上进行多目标蒸馏，本文采用了混合精度训练以加速蒸馏过程，参数见表 2。

表 2　多目标知识蒸馏训练参数

Tab. 2　Multi-objective knowledge distillation training parameters

参数	数值
蒸馏温度 T	3
交叉熵损失系数 α	5.0
掩码损失系数 β	2.0
余弦相似度损失系数 γ	1.0
学习率	2e-5

在支持多语言的释义识别对抗性数据集 paws-x 上，进行基于目标任务分数的参数剪枝，最终得到 Smaller-BERT，期以提升模型的泛化性和多语言适配性。

最后，本文采用了哈工大讯飞实验室开发的 Text-Pruner 工具包对模型进行剪枝，将隐层维度从原来的 3 072 降低到 2 048，将注意力头数量从原来的 12 个降低到 8 个。

以上就是 SmBERT 的训练全过程。

3.1　实验数据集及评价指标

分别在多任务自然语言理解基准和分析数据集 GLUE 的 QQP、QNLI、SST-2、MRPC、RTE 5 个数据

集上进行测试和评估。

QQP 是判断句对是否等价的二分类任务;QNLI 是判断句对是否蕴含的二分类任务;SST-2 是单句的情感二分类任务;MRPC 是判断句对是否释义等价的二分类任务;RTE 是识别文本蕴含的二分类任务。各个数据集样本分布见表 3。

表 3 实验各数据集样本分布

Tab. 3 Experimental sample distribution of each dataset /k

	QQP	QNLI	SST-2	MRPC	RTE
训练集	364	100	67	3.7	2.5
开发集	40	5.5	0.9	0.4	0.3
测试集	390	5.5	1.8	1.7	3

对于 SST-2,QNLI,RTE 数据集的评价指标为准确率(Precision),其计算公式为(15)

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

其中, TP 表示原样本为正例,预测结果也为正例的样本数, FP 表示原样本为负例,预测结果为正例的样本数。

召回率 Recall 指正类样本被找到的概率,式(16)。

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

其中, FN 表示原样本为正例,预测结果为负例的样本数。

QQP 和 MRPC 数据集的正样本比例过高,因此采用 $F1$ 值作为评价指标,用于衡量分类器性能的综合指标,其计算为公式(17)

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

3.2 实验结果与分析

为了验证 SmBERT 的压缩效果,本文选择了 BERT-base,GPT,BERT-large 在 5 个数据集的测试集上与 SmBERT 进行效果对比。鉴于设备的限制,模型的最长输入序列长度为 128,而不同的模型很可能采用不同的序列长度,因此为了更客观地探究 SmBERT 的性能,本文实验中除了 BERT-large 和 GPT,其余模型均是在同一设备下进行复现。具体的,各模型均采用 0.000 01 的学习率和 128 的最长序列长度,在相应任务数据集上进行微调,得出结果见表 4。

表 4 SmBERT 与不同模型在 5 个任务测试集的性能对比

Tab. 4 SmBERT performance comparison with different models in 5 task test sets

模型	参数量/M	QQP	QNLI	SST-2	MRPC	RTE	平均指标
BERT-base	110	71.6+	91.0+	93.9+	85.6-	64.4-	81.3
GPT	117	70.3	87.4	91.3	82.3	56.0	77.5
BERT-large	340	72.1	92.7	94.9	89.3	70.1	83.8
BERT-3	45.7	63.6	83.2	85.1	80.8	56.9	73.9
SmBERT	38.6	65.1	83.7	86.4	82.5	64.5	76.4

如表 4 所示,平均指标是模型在 5 个数据集上的平均分数,在多数据集比较中更能反映模型的泛化能力。对比平均指标可以看出,SmBERT 相比于原生的 BERT-base,只用了其 35%的参数量就达到了 94%的平均指标,具有较高压缩比,同时也维持了模型良好的多任务文本理解能力。当 SmBERT 与 GPT 模型对比时,发现 SmBERT 仅用了 GPT 模型 33% 的参数量就达到了其 99% 的平均性能。SmBERT 仅用其 11%的参数量就达到了其 91%的平均性能

此外,SmBERT 在 5 个任务上的表现均超越了 BERT-3(即 BERT 的前 3 层),而 SmBERT 的参数量也少于 BERT,证明了本文训练方法的有效性,要远优于直接在目标数据集上微调的子网络。

在 5 个数据集上 SmBERT 保持了 94%的 BERT-base 的语义能力,尤其是在小数据集 MRPC 表现了原模型 96%的能力,在 RTE 数据集上甚至超过了原模型的效果,说明 SmBERT 对小数据集任务更具优势。

为了探究模型对小数据集任务的性能和泛化能力,本文选择了与 RTE 样本规模最接近的 MRPC 任务的开发集上进行探究,实验结果见表 5。

表 5 各模型在 MRPC 开发集的 $F1$ 值

Tab. 5 The $F1$ value of each model in the MRPC development set

模型	层数	参数量	$F1$
BERT-base	12	110M	0.850
BERT-6	6	70.0M	0.836
DistilBERT	6	70.0M	0.836
Theseus	6	70.0 M	0.863
BERT-3	3	45.7M	0.762
DistilBERT-3	3	45.7M	0.753
Theseus-3	3	45.7M	0.789
SmBERT	3	38.6M	0.858

与 BERT, DistilBERT, Theseus 及其子网络对

比，SmBERT 在参数最少的情况下，在 MRPC 的开发集上的表现仅次于原生 Theseus，体现了模型在小数据集任务上良好的泛化性。同时 SmBERT 远超过同等深度的模型，体现了本文压缩方法的有效性。

为了验证 SmBERT 对推理速度的提升，本文分别应用 BERT-base 和 SmBERT 对维度为 32×512 的随机矩阵进行推理，推理时间对比见表 6。

表 6　推理时间对比

Tab. 6　Comparison of reasoning time

	平均推理时间/md	标准差/ms
BERT-base	831.83	5.58
BERT-6	540.04	2.47
BERT-3	362.30	2.52
SmBERT	136.14	1.15

从表 6 可以看出，SmBERT 的平均推理速度比 BERT-base 快了 6.1 倍，测试推理时间的标准差较小，说明模型的推理时间序列较为平滑，能反映模型正常推理速度。对比 BERT-3 和 SmBERT 可以发现，隐层和注意力头纵向的压缩对推理速度的提升也起到了一定的推理加速作用。

综上，SmBERT 只用了 BERT 参数量的 35%，在 5 个数据集上表现了 BERT 效果的 94%，并在小数据集任务上表现了更好的分类结果，推理速度提升了 6.1 倍。

4　结束语

本文提出了 SmBERT 的轻量型 BERT 类模型。首先，通过二次自蒸馏实现了模型的纵向压缩一倍；通过多目标知识蒸馏多维度地融合了大模型教师的语言知识；最后，基于跨语言任务的剪枝利用，降低了模型横向的参数冗余，提升泛化能力。

在与大模型和同类压缩模型的对比试验中，SmBERT 在高度压缩 BERT 模型的同时维持了良好的通用语言能力，更在小数据集上表现了良好的语言理解和迁移能力。

参考文献

[1] KOWSARI K, JAFARI MEIMANDI K, HEIDARYSAFA M, et al. Text Classification Algorithms: A Survey[J]. Information, 2019, 10(4):150.

[2] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China Technological Sciences, 2020, 63(10): 1872-1897.

[3] GUPTA M, AGRAWAL P. Compression of deep learning models for text: A survey [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2022, 16(4): 1-55.

[4] LAN Z, CHEN M, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.

[5] JIAO X, YIN Y, SHANG L, et al. Tinybert: Distilling bert for natural language understanding[J]. arXiv preprint arXiv: 1909.10351, 2019.

[6] SAJJAD H, DALVI F, DURRANI N, et al. On the effect of dropping layers of pre-trained transformer models[J]. arXiv preprint arXiv:2004.03844, 2020.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]. Advances in neural information processing systems, 2017:5998-6008.

[8] LECUN Y, DENKER J, SOLLA S. Optimal brain damage[J]. Advances in neural information processing systems, 1990: 598-605.

[9] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[10] MICHEL P, LEVY O, NEUBIG G. Are sixteen heads really better than one? [J]. arXiv preprint arXiv:1905.10650.2019.

[11] SUN Z, YU H, SONG X, et al. Mobilebert: a compact task-agnostic bert for resource-limited devices[J]. arXiv preprint arXiv:2004.02984, 2020.

(上接第 128 页)

[5] BZTA, BGDA, BMLA, et al. Combined electricity-heat-cooling-gas load forecasting model for integrated energy system based on multi-task learning and least square support vector machine[J]. Journal of Cleaner Production,2020,248:119252.

[6] 史佳琪,谭涛,郭经,等. 基于深度结构多任务学习的园区型综合能源系统多元负荷预测[J]. 电网技术,2018,42(3):698-707.

[7] WANG Xiaowei, LIU Wenjie, WANG Yingnan, et al. A hybrid NO_x emission prediction model based on CEEMDAN and AM-LSTM[J]. Fuel, 2021, 310: 122486.

[8] BEDI J, TOSHNIWAL D. Empirical mode decomposition based deep learning for electricity demand forecasting[J]. IEEE Access, 2018,6:49144-49156.

[9] 张学清,梁军,张熙,等. 基于样本熵和极端学习机的超短期风电功率组合预测模型[J]. 中国电机工程学报,2013,33(25):33-40,8.

[10] RICHMAN J S, MOORMAN J R. Physiological time-series analysis using approximate entropy and sample entropy [J]. American Journal of Physiology: Heart and Circulatory Physiology, 2000,278(6):2039-2049.

[11] HU Y, LI J, HONG M, et al. Short term electric load forecasting model and its verification for process industrial enterprises based on hybrid GA-PSO-BPNN algorithm—A case study of papermaking process[J]. Energy, 2019, 170: 1215-1227.

[12] Vaswani, Ashish, et al. Attention is all you need[C]// Proceeding of the 31st International Conference on Neural Information Processing Systems,2017:6000-6010.

[13] 邓带雨,李坚,张真源,等. 基于 EEMD-GRU-MLR 的短期电力负荷预测[J]. 电网技术,2020,44(2):593-602.