面向自然语言理解的多教师BERT模型蒸馏研究

石佳来，郭卫斌
华东理工大学信息科学与工程学院，上海 200237

摘要
知识蒸馏是一种常用于解决BERT等深度预训练模型规模大、推断慢等问题的模型压缩方案。采用“多教师蒸馏”的方法，可以进一步提高学生模型的表现，而传统的对教师模型中间层采用的“一对一”强制指定的策略会导致大部分的中间特征被舍弃。提出了一种“单层对多层”的映射方式，解决了知识蒸馏时中间层无法对齐的问题，帮助学生模型掌握教师模型中间层中的语法、指代等知识。在GLUE中的若干数据集的实验表明，学生模型在保留了教师模型平均推断准确率的93.9%的同时，只占用了教师模型平均参数规模的41.5%。

关键词
深度预训练模型；BERT；多教师蒸馏；自然语言理解

中图分类号：TP391.1 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023039

Multi-teacher distillation BERT model in NLU tasks

SHI Jialai, GUO Weibin
School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Abstract
Knowledge distillation is a model compression scheme commonly used to solve the problems of large scale and slow inference of BERT constant depth pre-training model. The method of "multi-teacher distillation" can further improve the performance of the student model, while the traditional "one-to-one" mapping method mandatory assignment strategy for the middle layer of the teacher model will lead to the abandonment of most of the middle features. The "one-to-many" mapping method is proposed to solve the problem that the middle layer cannot be aligned during knowledge distillation, and help students master the grammar, reference and other knowledge in the middle layer of the teacher model. Experiments on several data sets in GLUE show that the student model retains 93.9% of the average inference accuracy of the teacher model, while only accounting for 41.5% of the average parameter size of the teacher model.

Key words
deep pre-training model, BERT, multi-teacher distillation, nature language understanding

0 引言

知识蒸馏通常指包括规模较大的教师模型以及规模较小的学生模型，蒸馏时学生模型除了从样本的硬标签中学习外，还需要从教师模型的软标签中学习知识。Xie等人[1]于2021年提出的Elbert、Lan等人[2]于2019年提出的ALBERT、Jiao等人[3]于2019年提出的TinyBERT模型等针对BERT[4]的蒸馏模型都在自然语言处理（natural language processing，NLP）领域取得了较好的效果。最开始，学者们研究对BERT模型的蒸馏通常只采用单教师-单学生的架构，最近许多学者开始对多教师蒸馏模型进行研究。

对于不同的NLP任务，教师模型在假设空间、优化策略、参数初始化和许多其他因素方面有所不同，而表现出了不同的性能，因此，让学生模型从单一的教师模型中进行蒸馏学习，并不能满足多任务的场景。而采用多教师蒸馏策略时，虽然不同教师模型能为学生模型带来不同任务中的性能提升，但是提升效果并不明显，主要是因为在针对中间层知识的蒸馏时，教师模型的中间层数量是学生模型的数倍，传统的类似BERT-PKD的“单层对单层”的强制指定映射[5]的策略让学生模型学习到每个教师模型中间层知识十分有限，教师模型中的大部分的中间层知识难免会被舍弃，即使教师模型在某个特定任务上具有较好的性能，但是教师模型各自的优势被淡化了，学生模型即使想提取更多的知识表达也无能为力。为了解决该问题，本文提出了一种“单层对多层”的中间层映射方式，例如：当学生模型的中间层数为3，教师模型1的中间层数为12，教师模型2的中间层数为6时，学生模型的第1层中间层需要与教师模型1的第1、2、3层进行映射，也需要与教师模型2的第1、2层的中间层进行映射，并以此类推，计算并最小化中间层损失函数，从而使学生模型可以学习到每一个教师模型的每一个中间层的知识。

新的中间层蒸馏方法会为学生模型带来更大的学习成本，如何为学生模型寻找合适的教师模型也是一项值得研究的内容。对学生模型来说，教师模型并不是性能越强越好，若教师模型和学生模型之间的参数规模差距过大或教师模型非常复杂时，学生模型则有可能会无法接近教师模型[6]，因学生模型可能会无法捕捉复杂的教师模式捕获数据中更细粒度的知识，导致学生模型在数据的某些部分和某些其他部分过度拟合。为了解决上述问题，在固定$BERT_3$学生模型的情况下，分别对单个教师模型、两个教师模型、3个教师模型的蒸馏策略进行实验。在GLUE[7]任务中的QNLI、SST-2、MNLI数据集上进行了实验，验证了不同数量教师模型蒸馏出来的学生模型在自然语言推理任务、情感分类任务、释义相似性匹配任务方面的性能。

1 相关工作

1.1 预训练模型

预训练模型实质上属于一种迁移学习的策略，预训练模型先在原始任务上进行自监督训练，然后再针对特定的任务，对该初始模型进行微调，以获得在不同任务中的高效性能，使用者可以在自己的特定任务上使用别人训练好的模型。BERT模型是一种非常流行的预训练模型，采用

了Transformer[8]模型架构中的编码器模块，舍弃了解码器模块，由此组成的双向编码Transformer网络拥有强大的特征提取功能。

BERT模型通过掩码语言模型（mask language model，MLM）进行训练，在原始文本中随机抽取15%的分词进行掩码处理，即替换其中80%的分词为[MASK]标签，替换其中10%的分词为其他无关的分词，对其中10%的分词不做处理。预训练过程中BERT模型需要使用基于自注意力机制的全连接神经网络Transformer来表征上下文信息，Transformer层中主要包括两个部分：多头注意力（multi-head attention，MHA）和全连接前馈网络（feed forward network，FFN）。这样的结构可以让Transformer有一个显著的优势，即可以通过自注意力机制捕获输入分词之间的长距离依赖关系，这样的结构以CNN中的卷积思想为出发点，结合了多头注意力，在训练阶段实现了并行计算，帮助Transformer解决了传统神经网络算法训练耗时长的问题[9]。

基于BERT模型预训练的特点，不少学者将BERT模型运用在NLP的各种下游任务中，获得了不错的效果：李爱黎[10]等人曾在BERT模型基础上结合自扩展的中日文情感词典，提出了一个新的情感分析模型EmoBERT；韩立帆[11]等人在BERT模型的基础上添加了卷积层及句子级聚合等结构，以进一步优化生成的词表示，然后针对文言文标注数据稀缺的问题，构建了一个面向历史古籍文本标注任务的众包系统。

1.2 模型压缩与知识蒸馏

模型缩减的常见方案包括以下3类。

- Michel等人[12]发现Transformer中的注意力头虽然在训练时可以帮助模型很好地捕捉上下文信息，但是训练好的模型在预测阶段时，太多的注意力头部反而是冗余的，他们主张对训练好的模型的注意力头部进行剪枝，实验表明某些场景中一个注意力头部即可满足需求。
- Xu等人[13]把k-means聚类应用在卷积层，把权重矩阵分解为若干个小块，并引入了针对多个层级、单个层级的量化方法。Zafrir等人[14]提出了在模型的训练、微调阶段对模型的网络权重进行量化，在保持模型结构不变的同时，通过将模型权重量化为更小的数据类型。
- Hinton等人[15]认为教师模型庞大的参数可以帮助模型在训练阶段更好地掌握输入文本的知识表示，但是在预测阶段，过多的参数就显得有些冗余，并于2014年提出基于教师-学生架构的蒸馏模型。该模型通过知识蒸馏来训练更紧凑的学生模型，以接近教师模型的性能，让学生模型可以不经过教师模型相同规模的训练也可以拥有非常接近教师模型的表现，在许多领域中得到了很好的模型压缩的效果。

针对BERT模型的知识蒸馏方案可以分为：①从BERT模型蒸馏至长短期记忆网络（LSTM）等结构更加简单的模型；②从BERT模型蒸馏至同样拥有相同中间层结构，但是层级更浅的学生模型，或从BERT模型蒸馏至每个中间层更精简的学生模型，此类方案中教师模型与学生模型都拥有相似的Transformer中间层结构。

对于第一种蒸馏方案，Al-Omari等人[16]曾尝试选用双向长短记忆网络（bi-directional long short-term memory，BiLSTM）作为学生模型，可以大幅度地降低BERT教师模型的参数规模。杨秋勇等人[17]在此基础上提出了一种基于Bi-LSTM-CRF的模型，在中文特定领域（电力输送领域）的实体识别任务中获得了不错的效果。而叶榕等人[18]则在将BERT模型作为教师

模型的基础上，采用CNN为学生模型，在新闻文本分类领域取得了不错的效果。

对于第二种蒸馏方案，在知识蒸馏的过程中若采用更少模型层数的方案进行蒸馏，即采用BERT-to-BERT的蒸馏，可以更好地保留教师模型的性能。Xu等人[19]曾提出了BERT-of-Theseus，主张逐步将BERT中原始模块替换为更轻量级、参数更少的模块，在获得1.94倍预测速度提升的同时，只损失了1.4%的推断准确率。张睿东[20]将JointBERT模型作为初始模型，采用联合蒸馏的方法获得了在公共健康服务对话领域拥有高效性能的学生模型。Sun等人[5]于2019年提出了一种名为耐心蒸馏（patient knowledge distillation，PKD）的策略，旨在使学生模型除了从教师模型最后一层学习外，还让其从教师模型的中间层也学习知识[5]。Sun等人[5]还提出了跨层映射、尾层映射两种中间层映射的方法，如**图1**所示，“耐心蒸馏”额外在蒸馏损失函数中引入了归一化隐藏状态之间的均方差损失。Jiao等人[3]在此基础上提出的TinyBERT尝试对Transformer进行蒸馏，Jiao认为在Transformer蒸馏中，MHA层、FFN层的输出或一些中间表示（如注意力矩阵）可以作为损失函数中的一部分。本文同样在蒸馏过程中增加了对隐藏态、注意力矩阵的蒸馏以及对嵌入层的蒸馏。然而此类基于强制指定映射的蒸馏方案方法难以找到教师网络和学生网络之间的最

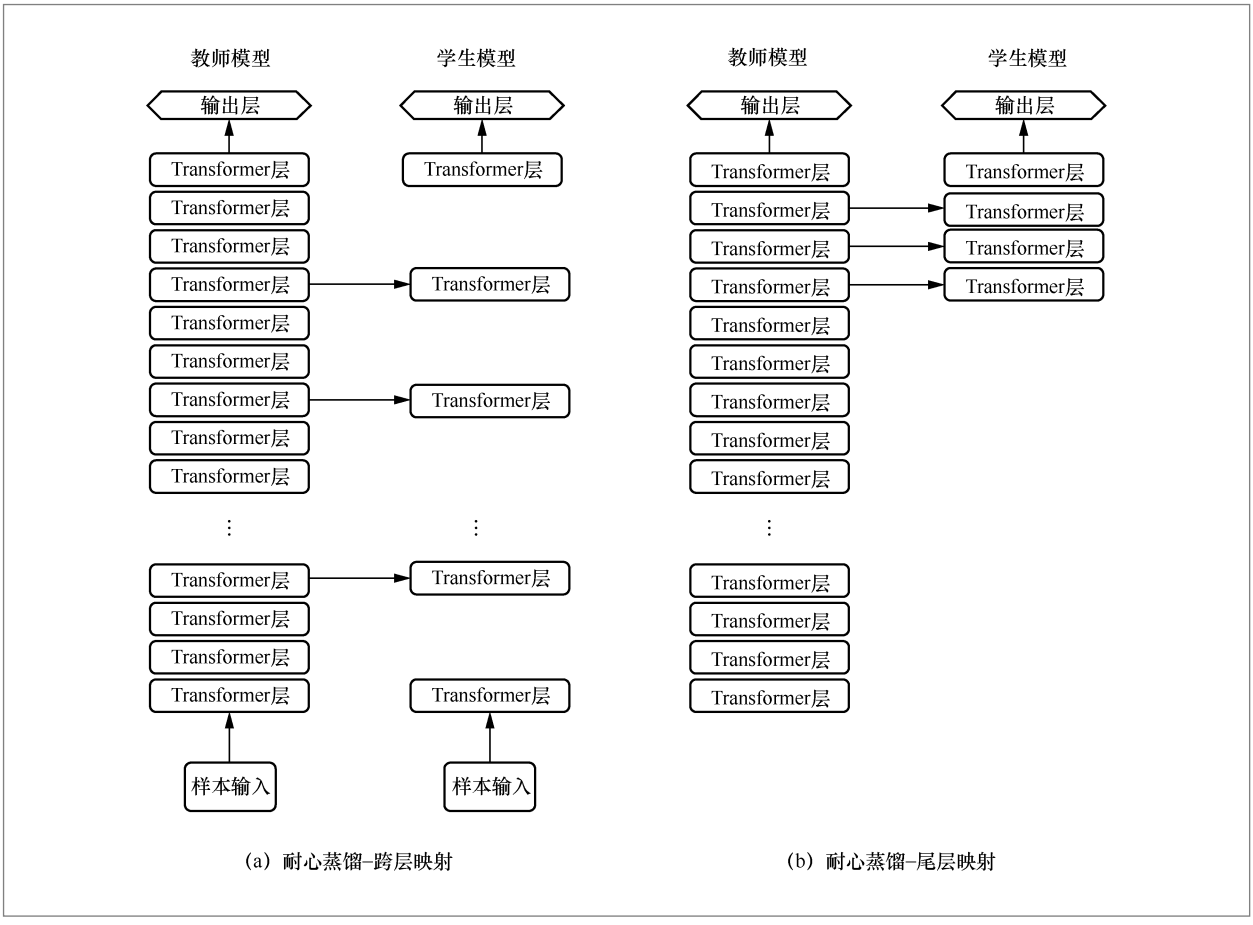

图1 耐心蒸馏的两种策略

佳层映射。让每个学生层仅从一个教师层学习，这可能会丢失教师网络中包含的丰富语言知识。

1.3 多教师蒸馏

Fukuda等人[21]于2017年提出的多教师蒸馏方案主张在蒸馏过程中直接使用多个教师数据。Cho等人[22]在2019年的研究中发现，在知识蒸馏过程中并不是性能越卓越的教师模型就一定可以蒸馏出更好的学生模型，这与人们的直观感受相悖，即只从教师模型的输出层学习知识并不能很有效地提升学生模型的表现。

多教师蒸馏的主要研究内容在于如何让多个教师模型在蒸馏过程中为学生模型提供更丰富的知识，即对多个教师模型的知识提取形式是一个较深刻的研究内容。为了更好地从多个教师模型中提取知识，Jiang等人[23]曾提出，同时使用两个教师模型训练学生模型，其中一个教师模型提供长期稳定的教师信息来保证与学生模型之间的差异，另一个教师模型提供短期的教师信息以提高学生模型学习的效率。同时Yang等人[24]于2020年提出了一种多教师两阶段蒸馏的方法，在多教师蒸馏的过程中同时按照Sun等人的“耐心蒸馏”的方案对每个教师模型的Transformer层进行蒸馏，针对机器问答任务取得了不错的表现。Wu等人[25]引入了共享池和预测层来对齐教师模型与学生模型之间的输出空间，从而更好地蒸馏，解决了当教师模型存在偏差时导致学生模型最终结果精度过低的问题。Yuan等人[26]为了解决在多教师蒸馏过程中，因每个教师模型都被分配了固定的权重而蒸馏效果不佳的问题，提出了一种利用增强学习的方法，为每个教师模型提供动态的权重。

多教师学习是通过利用多个教师模型提高学生模型在单个任务上的性能。多教师蒸馏方法的核心在于让学生模型从多个教师软标签、中间层中获取知识，以上针对多教师中间层知识的获取方案仅仅停留在对教师模型少数中间层的知识提取，学生模型因为无法完整地获取每个教师模型所提供的“观点”而并不能很好地针对每个教师模型的优势进行学习。本文提出的多教师蒸馏方案，让学生模型到每个教师模型中均获取整体的中间层的知识。

2 多教师蒸馏模型介绍

2.1 模型整体架构

本文提出的多教师蒸馏模型的整体架构如**图2**所示，由多个预训练好的教师模型同时蒸馏单个学生模型。与Jiao等人[3]提出的TinyBERT相似的是，本实验的蒸馏损失函数包括嵌入层蒸馏损失函数、Transformer层的蒸馏损失函数和预测层蒸馏损失函数。其中对于多层的Transformer，采用“单层对多层”的方式进行映射。Transformer层的蒸馏是基于注意力和隐藏状态的蒸馏，每个学生注意力层/隐藏层可以从每个教师模型的多个注意力层/隐藏层中学习知识。

2.2 预测层损失函数

学生模型和教师模型都会在每个样本上产生一个关于类别标签的分布信息，软标签损失函数的计算就是计算这两个分布之间的相似性。假设实验使用K个教师模型进行实验，需要计算学生模型预测层输出和K个教师模型预测输出的交叉熵（CE），如式（1）所示（以下式中均用CE代替）。

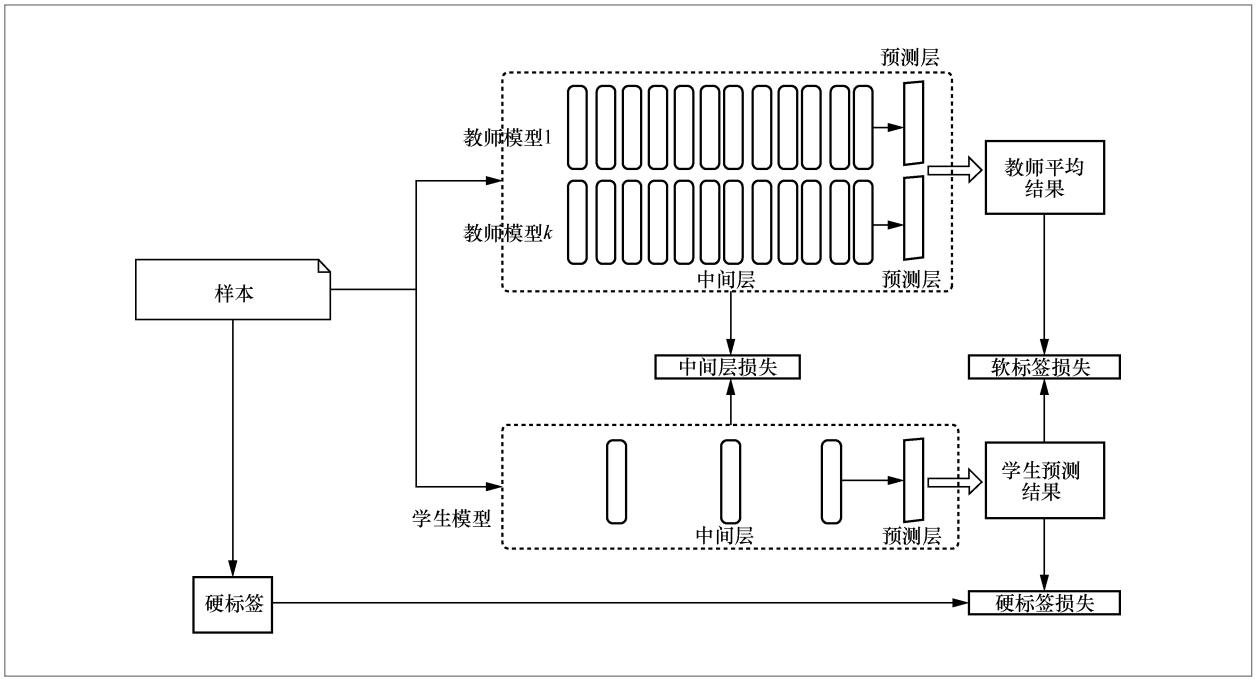

图 2 提取中间层知识的多教师模型蒸馏模型

$$\mathcal{L}_{\text{predict}} = \frac{1}{K}\sum_{k=1}^{K}\text{CE}\left(\text{softmax}\left(\frac{z_k^T}{\tau}\right),\text{softmax}\left(\frac{z^S}{\tau}\right)\right) = -\frac{1}{K}\sum_{k=1}^{K}\left(\text{softmax}\left(\frac{z_k^T}{\tau}\right)\right)\cdot\log\left(\text{softmax}\left(\frac{z^S}{\tau}\right)\right) \tag{1}$$

其中，z_k^T 表示第k个教师模型的预测的对数概率值，$\frac{z^S}{\tau}$ 表示学生模型预测的对数概率值，使用 softmax 函数将对数概率值 z_i映射到概率向量$\boldsymbol{p}_i$中，这样的映射可以使其每个映射的值和为1。τ表示蒸馏时的温度参数，温度越高，概率分布曲线越“平滑”，即淡化各个标签之间预测值的差异。

2.3 隐藏层损失函数

针对Transformer层的蒸馏包括对FFN的隐藏层的蒸馏和对注意力层的蒸馏[25]。学生和教师模型Transformer层之间的映射如**图3**所示，对比Sun等人[5]提出的“耐心蒸馏”的指定跨层映射，本文的“单层对多层”的蒸馏方案可以完整地获取每个教师模型的所有中间层信息，在多教师蒸馏中更好地获取每个教师模型的“看法”。

在计算隐藏态损失函数时，假设共有K个教师模型，设第k个教师模型拥有M层Transformer，学生模型拥有N层Transformer，那么需要将第k个教师模型的中间层按顺序平均分为N组，第l层学生模型将学习第l组教师模型中间层中的隐藏态、注意力矩阵的知识。对于学生模型的第i层中间层，需要学习第k个教师模型的第i组的中间层，本文将第k个教师模型第i组的中间层表示为 $I_{k,i}$ 。以第一个教师模型BERT_{12}和学生模型BERT_4为例，对于第1组，$I_{1,1}=\{1,2,3\}$ 。学生模型第i层的隐藏态可以表示为 h_i^S ，第k个教师模型的第j层的

隐藏态可以表示为 $\boldsymbol{h}_j^{\mathrm{T},k}$ 。使用 $W_{h,k}$ 代表一个线性变换参数，将学生的隐藏态转换为与教师网络状态相同的空间学生模型第i层中间层和第k个教师模型第j层中间层之间的隐藏态距离可以由式(2)表示：

$$d_{i,j}^{h,k}=\mathrm{MSE}(\boldsymbol{h}_i^{S}W_{h,k},\boldsymbol{h}_j^{\mathrm{T},k}) \tag{2}$$

若假设第i层学生模型映射的第k个教师模型的中间层起始位置是p，结束位置是q，则学生模型第i层中间层隐藏态到教师模型第i组中的各个中间层隐藏态之间的加权平均距离记为其到某一组教师模型中间层的距离，如式(3)所示：

$$\hat{d}_i^{h,k}=\frac{1}{q-p}\sum_{j=p}^{q}d_{i,j}^{h,k} \tag{3}$$

将学生模型所有中间层隐藏态到第k个教师模型对应的中间层隐藏态组之间的距离之和记为学生模型到第k个教师模型整体的隐藏态目标函数，如式(4)所示：

$$D_k^h=\sum_{i=1}^{N}\hat{d}_i^{h,k} \tag{4}$$

学生模型对每个教师模型均需要按如上的方式进行映射，以计算每个教师模型的层隐藏态的距离，再将学生模型隐藏态到每个教师模型的隐藏态的距离进行求和，即可得到整个多教师蒸馏过程中的隐藏态损失函数，如式(5)所示：

$$\mathcal{L}_{\mathrm{hidden}}=\frac{1}{K}\sum_{k=1}^{K}D_k^h \tag{5}$$

2.4 注意力层损失函数

Clark等人[27]于2019年提出BERT学习的注意力权重可以表示丰富的语言信息知识，这种语言知识包括语法信息和指代信息，这对于自然语言的理解是必不可少的。具体做法与Jiao等人[3]提出的TinyBERT相似，让学生模型学习适应教师模型中的多头注意力矩阵。注意力矩阵的计算方式如式(6)所示。其中δ是键的维度，作为缩放因子；$\boldsymbol{a}$是通过点积运算从Q和K的相容性计算出的注意力矩阵。

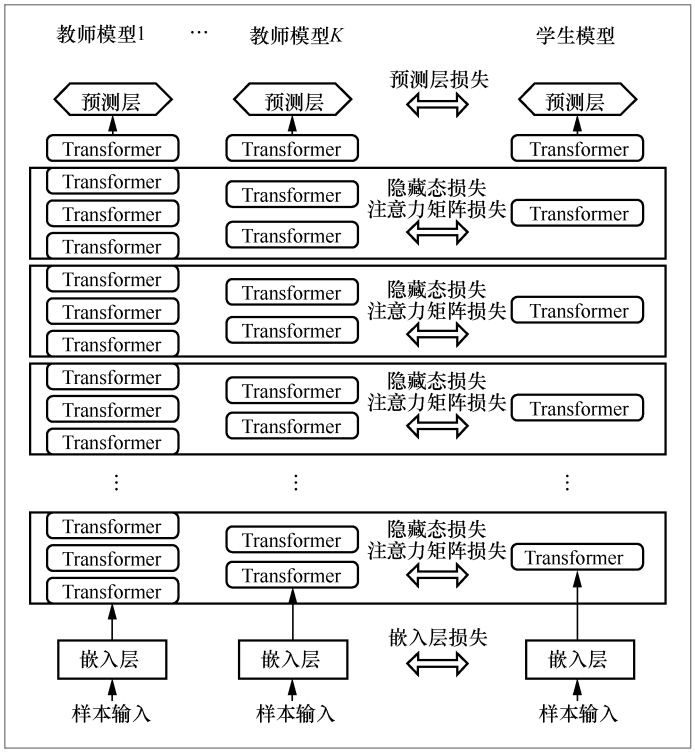

图 3 多教师蒸馏中的跨层映射策略

$$\boldsymbol{a}=\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{\delta}} \tag{6}$$

与隐藏态损失计算相似，第i层学生模型的注意力层需要分别从K个教师模型中学习 $I_{k,i}$ 组的中间层的注意力矩阵的知识。第i层学生模型的注意力矩阵可以表示为 $\boldsymbol{a}_i^{S}$ ，同样地，本文也用 $\boldsymbol{a}_j^{\mathrm{T},k}$ 表示第k个教师模型的第j层教师模型的注意力矩阵，假设 h_{num} 是注意力头部的数量。学生模型第i层中间层和第k个教师模型第j层中间层之间的注意力矩阵的距离可以由式(7)表示：

$$d_{i,j}^{a_k}=\frac{1}{h_{\mathrm{num}}}\sum_{h=1}^{h_{\mathrm{num}}}\mathrm{MSE}(\boldsymbol{a}_{i,h}^{S},\boldsymbol{a}_{j,h}^{\mathrm{T},k}) \tag{7}$$

与隐藏态计算相似，假设第i层学生模型映射的第k个教师模型的中间层起始位

置是p，结束位置是q，则学生模型第i层中间层注意力矩阵到教师模型第i组中的各个中间层注意力矩阵之间的加权平均距离记为其到某一组教师模型中间层的注意力矩阵距离，如式(8)所示：

$$\hat{d}_i^{a,k}=\frac{1}{q-p}\sum_{j=p}^{q}d_{i,j}^{a,k} \tag{8}$$

将学生模型所有中间层注意力矩阵到第k个教师模型对应的中间层注意力矩阵组之间的距离之和记为学生模型到第k个教师模型整体的注意力矩阵目标函数，如式(9)所示：

$$D_k^a=\sum_{i=1}^{N}\hat{d}_i^{h,k} \tag{9}$$

学生模型对每个教师模型均需要按如上方式进行映射以计算到每个教师模型的层注意力矩阵的距离，再将学生模型注意力矩阵到每个教师模型的注意力矩阵的距离求和即可得到整个多教师蒸馏过程中的注意力矩阵损失函数，如式(10)所示：

$$\mathcal{L}_{\text{attention}}=\frac{1}{K}\sum_{k=1}^{K}D_k^a \tag{10}$$

2.5 嵌入层损失函数

词嵌入是NLP中语言模型与表征学习技术的统称，它是指把包括所有词数量的高维空间嵌入一个低维连续向量空间中，每个单词或词组被映射为实数域上的向量。若希望更好地表示单词，需要花费更多的内存空间，在不牺牲模型性能的情况下压缩嵌入矩阵对于实际应用至关重要。为此使用均方误差（MSE）来表示教师模型和学生模型词嵌入之间的“距离”，并通过最小化它来提升学生模型性能。第k个教师模型的词嵌入为$\mathrm{E}^{\mathrm{T},k}$，$\mathrm{E}^{S}$表示具有相同形状的学生模型的词嵌入，使用$W_{e,k}$代表一个线性变换参数，将学生的隐藏态转换为与教师网络状态相同的空间。嵌入层的损失函数如式(11)所示：

$$\mathcal{L}_{\text{embdding}}=-\frac{1}{K}\sum_{k=1}^{k=K}\mathrm{MSE}(\mathrm{E}^{S}W_{e,k},\mathrm{E}^{\mathrm{T},k}) \tag{11}$$

2.6 整体损失函数

本文将预测层损失、隐藏层损失、注意力层损失、嵌入层损失以线性的方式整合在一起，得出最后的整体蒸馏损失函数，如式(12)所示：

$$\begin{aligned}\mathcal{L}_{\text{distill}}&=\alpha\mathcal{L}_{\text{hidden}}+\mathcal{L}_{\text{attention}}+\\&\mathcal{L}_{\text{embedding}}+(1-\alpha)(\mathcal{L}_{\text{prediction}})\end{aligned} \tag{12}$$

蒸馏损失函数结合学生模型与样本硬标签之间的损失函数可以得到在整个蒸馏过程中的损失函数，如式(13)所示：

$$\mathcal{L}=\beta\mathcal{L}_{\text{distill}}+(1-\beta)(\mathcal{L}_{\text{CE}}) \tag{13}$$

3 实验实施与结果

3.1 实验设置失函数

本实验采用GLUE[7]中的公开数据集QNLI、MNLI、SST-2进行实验，以验证模型在该自然语言理解任务中的性能。对于QNLI数据集，模型需要判断问题和句子是否为蕴含关系，结果包括蕴含和不蕴含两种情况，属于二分类问题。QNLI包括训练集104 743个、开发集5 463个、测试集5 461个。对于SST-2数据集，模型需要判断文本蕴含的情感是正面情绪还是负面情绪，也是二分类问题，包括训练集67 350个、开发集873个、测试集1 821个。对于MNLI，模型需要根据给定的前提语句和假设语句，预测前提语句是否包含假设、与假设矛盾，或者两者都不，属于三分类问题。MNLI

集合了许多不同领域风格的文本，根据文本内容是否相匹配分为匹配和不匹配两个版本的数据集，前者指的是训练集和测试集的数据来源一致，后者则表示不一致。本文选用匹配版本的数据集，包括训练集392 702个，匹配版的开发集9 815个，匹配版的测试集9 796个。

为了方便蒸馏时进行跨层映射，选用的教师模型都为12层基于BERT的预训练好的模型，分别选用$BERT_{12}$、$RoBERTa_{12}$[28]、$XLNet_{12}$[29]作为教师模型。教师模型选择的原因包括：①以上教师模型中间层均采用Transformer结构，有利于进行中间层的知识蒸馏；②以上3个教师模型都是公认效果较好的BERT模型的变体，结构相似；③以上3个教师模型被许多多教师蒸馏相关的文献（Yuan等人[26]、Yang等人[24]）作为教师模型。每个教师模型均包括1.1亿个参数。每个数据集都采用不同的随机种子微调了3个教师模型。各个教师模型在每个数据集上的表现见**表1**，各个数据集中均用准确率表示结果。不同的教师模型在不同的任务中表现不同，不存在一个模型在所有任务上比其他模型表现得都要好，这是因为业务不同的模型可能学习到不同的局部最优，并且每个模型承受的偏差也不同。在自然语言推断、情感分类任务上，$RoBERTa_{12}$教师模型的性能表现最出色且优势最明显，这是因为$RoBERTa_{12}$模型在预训练阶段采用了更多数据集进行训练。而在释义性分析任务中，$XLNET_{12}$教师模型的表现则最突出。

选取的学生模型为$BERT_3$，先使用$BERT_{12}$的前3层的参数作为初始值，再用前文的知识蒸馏框架对学生模型进行预训练。学生模型拥有0.45亿个参数。为了验证教师模型的数量与蒸馏的关系，本实验分别设立了单个教师模型、两个教师模型、3个教师模型的蒸馏组别。

表1 各个教师模型的表现

教师模型	QNLI	SST-2	MNLI-m	平均值
$BERT_{12}$	91.1%	91.3%	83.8%	88.7%
$RoBERTa_{12}$	91.8%	93.5%	84.3%	89.8%
$XLNET_{12}$	90.9%	92.7%	85.8%	89.8%

其中，单个教师的对照组分别设立单独选取{$BERT_{12}$}、{$RoBERTa_{12}$}、{$XLNET_{12}$}为教师模型的对照，两个教师模型的对照组分别设立选取{$BERT_{12}$ + $RoBERTa_{12}$}、{$BERT_{12}$ + $XLNET_{12}$}、{$RoBERTa_{12}$+$XLNET_{12}$}为教师模型的对照组；3个教师模型的对照组设立选取{$BERT_{12}$ + $RoBERTa_{12}$ + $XLNET_{12}$}为教师模型的对照组。

训练时式(1)中的蒸馏温度设置为{1,5,10}，学习率设置为{1×10^{5},2×10^{5},5×10^{-5}}，式(12)中的权重设置为{0.1,0.2,0.5}。式(13)中用于平衡教师提供的软标签的所有损失和真实类别标签的损失的参数β设置为{0.2,0.5,0.7}。批量的大小设置为32，最多对数据进行4轮训练。

3.2 不同教师数量蒸馏对比

为了获取在本文实验环境中性能更好的教师模型的组合，对不同教师数量的蒸馏实验中，学生模型的性能见**表2**。结果显示选用{$BERT_{12}$ + $RoBERTa_{12}$}教师模型组的蒸馏性效果最佳，两个教师模型的平均推断准确率为89.3%，而学生模型的平均推断准确率为83.9%。学生模型的推断准确率在保留了各个教师模型平均推断准确率的93.9%的同时，参数规模只占用了教师模型平均参数规模的41.5%。

首先，可以发现在蒸馏时，学生模型可以继承教师模型在某一特定任务中的优秀性能，例如在各个教师模型中，RoBER-

表 2 在不同教师数量情况下蒸馏获得的学生模型性能

教师数量	教师模型	QNLI	SST-2	MNLI-*m*	平均值
1	$BERT_{12}$	84.1%	85.1%	74.4%	81.2%
1	$RoBERTa_{12}$	84.3%	86.3%	75.3%	81.9%
1	$XLNET_{12}$	84.5%	85.4%	71.3%	80.4%
2	$BERT_{12}$+ $RoBERTa_{12}$	85.8%	89.2%	76.8%	83.9%
2	$BERT_{12}$+ $XLNET_{12}$	85.7%	88.4%	76.3%	83.4%
2	$RoBERTa_{12}$+ $XLNET_{12}$	85.6%	89.1%	75.7%	83.4%
3	$BERT_{12}$+ $RoBERTa_{12}$+ $XLNET_{12}$	85.7%	88.6%	76.8%	83.7%

Ta_{12}教师模型在SST-2的情感分类任务上的性能是最佳的，结果见**表1**，在这一前提下，横向来看，在单教师蒸馏组别中，选用{$RoBERTa_{12}$}教师模型的组别比其他两个单教师蒸馏组别中的学生模型在该任务中的性能更加卓越，准确率达到了86.3%。纵向来看，在全部数量组别中，在该任务中表现最好的蒸馏模型是采用{$BERT_{12}$ + $RoEBRTa_{12}$}两个教师模型的组别，准确率达到了89.2%。

最后，还可以发现更强的教师模型不一定可以蒸馏出更好的学生模型，同时更多的教师模型也不一定可以蒸馏出更好的学生模型。例如采用{$BERT_{12}$+ $RoBERTa_{12}$+ $XLNET_{12}$}3个教师模型的蒸馏组别中获得的学生模型并没有在自然语言推断、情感分类任务中优于采用{$BERT_{12}$+$RoBERTa_{12+}$}两个教师模型的蒸馏组别。

3.3 中间层蒸馏效果验证

为了验证本文提出的蒸馏方案对多教师蒸馏性能的提升，本实验还与其他基线模型进行了对比。同时还分别在单个教师、两个教师、3个教师组别中选取了在该组别中性能最好的蒸馏方案，分别是：{$RoBERTa_{12}$}、{$BERT_{12}$ + $RoBERTa_{12}$}和{$BERT_{12}$ + $RoBERTa_{12}$ + $XLNET_{12}$}。本文采用的蒸馏方案记作Transformer层蒸馏（transformer knowledge distillation，TKD）。再按所选取的蒸馏组别，在分别设立教师模型相同、损失函数不同的两个对照组。新设立的对照组1均只从教师预测层中学习知识，此蒸馏策略记为软标签蒸馏（original knowledge distillation，OKD），新设立的对照组2的蒸馏策略与Sun等人的“耐心蒸馏”[5]一致，从指定的教师中间层中学习知识，此蒸馏策略记为耐心蒸馏（PKD），实验结果见**表3**。

从横向进行比较，可以看到在固定教师模型的数量、类型相同的情况下，本文提出的蒸馏方法对Transformer层的知识蒸馏在结果的平均值上均为最优，除了在3个教师模型的SST-2数据集中出现了意外，其余每个数据集中几乎均是本文蒸馏方案的结果更好，证明了本文提出的蒸馏方案可以一定程度上提升学生模型的性能。本文提出的Transformer层蒸馏方案对比软标签蒸馏蒸馏方案提升空间较大，对比耐心蒸馏方案提升空间虽不是特别明显，但是几乎在各个任务、教师模型数量的情况下也均有提升。例如在同选用{$BERT_{12}$ + $RoBERTa_{12}$}两个教师模型的情况下，本文的耐心蒸馏策略比另外两种基线模型的性能要好，对比软标签蒸馏方案差别最大的是在MNLI-*m*数据集中，准确率提升了3.4%；对比耐心蒸馏方案差别最大的是在QNLI数据集中，准确率提升了1.1%；而在平均值上

表3 在不同教师数量情况下蒸馏获得的学生模型性能

教师数量	蒸馏策略	QNLI	SST-2	MNLI-*m*	平均值
1	针对Transformer层的蒸馏	84.3%	86.3%	75.3%	81.9%
1	耐心蒸馏	83.2%	86.1%	75.0%	81.4%
1	软标签蒸馏	82.1%	85.1%	71.9%	79.7%
2	针对Transformer层的蒸馏	85.8%	89.2%	76.8%	83.9%
2	耐心蒸馏	85.0%	88.1%	74.1%	82.4%
2	软标签蒸馏	83.1%	86.3%	70.4%	79.9%
3	针对Transformer层的蒸馏	85.7%	88.6%	76.8%	83.7%
3	耐心蒸馏	84.8%	89.0%	76.3%	83.3%
3	软标签蒸馏	82.9%	87.2%	72.7%	80.9%

对比差别最大的OKD提升了2.2%；

但是，这种现象在引入更多教师模型的蒸馏情况中就显得不是那么明显了，例如在引入{$BERT_{12}$+$RoBERTa_{12}$+$XLNET_{12}$}3个教师模型的蒸馏情况中，虽然平均值还是领先于另外两种蒸馏策略，但是在SST-2数据集中，出现了本文的Transformer层蒸馏策略被反超的情况，尽管只被耐心蒸馏超了0.4%。这里可以理解为更多的教师模型已经为学生模型提供了非常丰富的知识，再加上学生模型和教师模型之间的参数规模比较大，$BERT_3$学生模型捕捉教师模型中间知识的性能并不是很好。

3.4 与现有蒸馏模型对比

为了进一步验证本文提出的蒸馏方案对模型性能的提升，本实验还分别与多个现有的多教师蒸馏模型、单教师蒸馏模型进行了对比。其中多教师蒸馏组中均选用{$BERT_{12}$ + $RoBERTa_{12}$}的组合作为教师模型。单教师蒸馏与多教师蒸馏对照组中的学生模型均选用$BERT_3$。

针对多教师蒸馏的对照组，本实验选用如下基线模型。

- RL-KD（reinforced learning based knowledge distillation）：Yuan等人[26]提出的使用强化学习在多教师蒸馏实验中为每个教师模型计算动态权重的多教师蒸馏模型。
- RSE-KD（random-single-ensemble teacher knowledge distillation）：Fukuda等人[21]提出的随机从小批量教师模型中选一名教师为学生模型提供软标签的多教师蒸馏模型。

针对单教师模型的对照组，本实验选用如下基线模型。

- PKD_3：Sun等人[5]提出的“耐心”蒸馏模型，即将教师模型与学生模型的中间层进行强制指定的蒸馏模型。
- $TinyBERT_3$：Jiao等人[3]提出的在预训练阶段、微调阶段都进行蒸馏的针对Transformer-based模型的蒸馏模型。

实验结果见**表4**，可以看到在相同学生模型的条件下，无论是单教师蒸馏对照组还是多教师蒸馏对照组，均是本文提出的方法更优，这验证了本文提出的针对Transformer蒸馏的多教师模型可以帮助学生模型从每一个教师模型的全部中间层中获取知识，并且知识的获取可以真实地帮助学生模型提升模型的性能。本文模型的平均准确率提升在单教师模型中更明显，证明本文模型可以让多个教师模型更好地发挥其优势，在蒸馏实验中对学生模型的性能提升更大。

表 4 与现有蒸馏模型对比

蒸馏方案	模型	QNLI	SST-2	MNLI-*m*	平均值
单教师	$TinyBERT_3$[3]	83.8%	86.6%	75.4%	81.9%
单教师	PKD_3[5]	83.2%	86.1%	75.0%	81.4%
多教师	RL-KD[26]	84.3%	88.5%	76.7%	83.1%
多教师	RSE-KD[21]	83.7%	86.0%	75.0%	81.5%
多教师	TKD(ours)	85.8%	89.2%	76.8%	83.9%

4 结束语

本文针对传统多教师蒸馏只蒸馏教师模型预测层而忽略中间层之上表达的问题，提出了对BERT模型的多教师蒸馏方法，同时修改了传统的蒸馏损失函数，新增了对中间Transformer层的知识的提取。实验选用预训练好的$BERT_{12}$、$RoBERTa_{12}$、$XLNET_{12}$作为教师模型，学生模型统一采用$BERT_3$模型。再分别按排列组合设立单个教师、两个教师、3个教师的实验组，验证教师模型数量对蒸馏实验的影响。实验结果表明，采用{$BERT_{12}$ + $RoBERTa_{12}$}教师模型组的蒸馏实验效果最佳。

为了验证对Transformer的蒸馏对多教师蒸馏的影响，实验还分别设立了只从教师模型预测层学习、传统耐心蒸馏的单个教师、两个教师、3个教师的蒸馏对照组进行实验。最后发现在两个教师模型的蒸馏对照试验中，Transformer的蒸馏可以明显提升学生模型性能的提升。而这种现象会在更多教师模型的蒸馏实验中（3个教师模型）减弱，这是因为规模较小的学生模型难以捕捉过于丰富的教师模型中间层信息。

综上所述，协调教师模型数量、教师模型选取以及对教师模型的蒸馏方案，是获取最适合的多教师蒸馏方案的关键，同时新的蒸馏损失函数可以帮助学生模型更好地从多个教师模型的中间层学习知识。

参考文献:

[1] XIE K, LU S, WANG M, et al. Elbert: fast albert with confidence-window based early exit[C]//Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2021: 7713-7717.

[2] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[C]// Proceedings of 8th International Conference on Learning Representations. New York: OpenReview.net, 2020: 564-571.

[3] JIAO X Q, YIN Y C, SHANG L F, et al. TinyBERT: distilling BERT for natural language understanding[C]//Proceedings of the Association for Computational Linguistics. New York: EMNLP, 2020: 4163-4174.

[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter ofthe Association for Computational Linguistics: Human Language Technologies. [S.l.]: ACL Press, 2019: 4171-4186.

[5] SUN S Q, CHENG Y, GEN Z, et al. Patient knowledge distillation for BERT model compression[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language, Processing and the 9th International Joint Conference on Natural Language Processing. New York:

EMNLP-IJCNLP, 2019: 4322-4331.

[6] ILICHEV A, SOROKIN N, PIONTKOVSKAYA I, et al. Multiple teacher distillation for robust and greener models[C]//Proceedings of the International Conference on Recent Advances in Natural Language Processing. New York: RANLP, 2021: 601-610.

[7] WANG A, SINGH A, MICHAEL J, et al. GLUE: a multi-task benchmark and analysis platform for natural language understanding[J]. arXiv preprint, 2018, arXiv:1804.07461.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30(1): 5998-6008.

[9] 任欢, 王旭光. 注意力机制综述[J]. 计算机应用, 2021, 41(z1): 1-6.
REN H, WANG X G. Overview of attention mechanism[J]. Computer Applications, 2021, 41(z1): 1-6.

[10] 李爱黎, 张子帅, 林荫, 等. 基于社交网络大数据的民众情感监测研究[J]. 大数据, 2022, 8(6): 105-126.
LI A L, ZHANG Z S, LIN Y, et al. Research on public emotion monitoring based on social network big data[J].Big Data Research, 2022, 8(6): 105-126.

[11] 韩立帆, 季紫荆, 陈子睿, 等. 数字人文视域下面向历史古籍的信息抽取方法研究[J]. 大数据, 2022, 8(6): 26-39.
HAN L F, JI Z J, CHEN Z R, et al. Research on information extraction from historical ancient books from the perspective of digital humanities[J]. Big Data Research, 2022, 8(6): 26-39.

[12] MICHEL P, LEVY O, NEUBIG G. Are sixteen heads really better than one?[J]. Advances in Neural Information Processing Systems, 2019, 32(1): 4809-4818.

[13] XU Y, WANG Y, ZHOU A, et a1. Deep neural network compression with single and multiple level quantization[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New York: ACM Press, 2018.

[14] ZAFRIR O, BOUDOUKH G, IZSAK P, et al. Q8bert: quantized 8bit bert[C]// Proceedings of 2019 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS). Piscataway: IEEE Press, 2019: 36-39.

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

[16] Al-OMARI H, ABDULLAH M A, SHAIKH S. Emodet2: emotion detection in english textual dialogue using BERT and BiLSTM models[C]//Proceedings of 2020 11th International Conference on Information and Communication Systems. Piscataway: IEEE Press, 2020: 226-232.

[17] 杨秋勇, 彭泽武, 苏华权, 等. 基于Bi-LSTM-CRF的中文电力实体识别[J]. 信息技术, 2021(9): 45-50.
YANG Q Y, PENG Z W, SU H Q, et al. Chinese power entity recognition based on Bi-LSTM-CRF[J]. Information Technology, 2021(9): 45-50.

[18] 叶榕, 邵剑飞, 张小为, 等. 基于BERT-CNN的新闻文本分类的知识蒸馏方法研究[J]. 电子技术应用, 2023, 49(1): 8-13.
YE R, SHAO J F, ZHANG X W, et al. Research on knowledge distillation method of news text classification based on BERT-CNN [J] Application of Electronic Technology, 2023, 49(1): 8-13.

[19] XU C, ZHOU W, GE T, et al. BERT-of-theseus: compressing BERT by progressive module replacing[C]// Proceedings of Empirical Methods in Natural Language Processing. 2021: 7859-7869,

[20] 张睿东. 基于BERT和知识蒸馏的自然语言理解研究[D]. 南京: 南京大学, 2020.
ZHANG R D. Research on natural language understanding based on BERT and knowledge distillation[D]. Nanjing: Nanjing University, 2020.

[21] FUKUDA T, KURATA G. Generalized knowledge distillation from an ensemble of specialized teachers leveraging

unsupervised neural clustering[C]// Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6868-6872.
[22] CHO J H, HARIHARAN B. On the efficacy of knowledge distillation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.:s.n.], 2019: 4794-4802.
[23] JIANG L, WEN Z, LIANG Z, el al. Long short-term sample distillation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM Press, 2020: 4345-4352.
[24] YANG Z, SHOU L, GONG M, et al. Model compression with two-stage multi-teacher knowledge distillation for web question answering system[C]// Proceedings of the 13th International Conference on Web Search and Data Mining. [S.l.:s.n.], 2020: 690-698.
[25] WU C, WU F Z, HUANG Y F. One teacher is enough? Pre-trained language model distillation from multiple teachers[C]// Proceedings of the Association for Computational Linguistics. New York: ACL Press, 2021: 4408-4413.
[26] YUAN F, SHOU L, PEI J, et al. Reinforced multi-teacher selection for knowledge distillation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(16):14284-14291.
[27] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: pre-training text encoders as discriminators rather than generators[C]// Proceedings of 8th International Conference on Learning Representations. New York: ICLR, 2020.
[28] LIU Z, LIN W, SHI Y, et al. A robustly optimized BERT pre-training approach with post-training[C]//Proceedings of Chinese Computational Linguistics: 20th China National Conference. Cham: Springer, 2021: 471-484.
[29] YANG Z L, DAI Z L, CARBONELL J G, et al. XLNet: generalized autoregressive pretraining for language understanding[C]// Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. New York: NeurIPS, 2019: 5754-5764.

作者简介

石佳来(1998-),男,华东理工大学信息科学与工程学院硕士生,主要研究方向为自然语言理解、知识蒸馏等。

郭卫斌(1968-),男,博士,华东理工大学信息科学与工程学院教授,中国计算机学会高级会员,主要研究方向为高性能计算、大数据与云计算、计算机应用等。

收稿日期: 2023-01-16
基金项目: 国家自然科学基金项目(No. 62076094)
Foundation Item: The National Natural Science Foundation of China (No. 62076094)