Computer Science and Application 计算机科学与应用, 2023, 13(8), 1515-1526
Published Online August 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.138150

Hans汉斯

基于多教师知识蒸馏的新闻文本分类方法

杜潇鉴*，吕卫东#，孙钰华

兰州交通大学数理学院，甘肃 兰州

收稿日期：2023年7月5日；录用日期：2023年8月3日；发布日期：2023年8月14日

摘　要

从传统的文本分类到基于深度学习下的文本分类，再到BERT模型的提出，使得其以及其变种模型逐渐成为自然语言处理中的主流模型，但其需要占用和花费大量内存和计算机资源。根据师生网络结构分成同构和异构两种情况，并提出了不同的多教师蒸馏策略。在THUCNews数据集上做实验，发现即使有教师表现较差，也能使得学生模型分类效果分别提升3.26%和3.30%，且性能损失分别为0.79%和0.78%，说明接近教师模型的分类效果；同时参数量只是教师模型的2.05%和2.08%，实现了很好的模型压缩。

关键词

知识蒸馏，多教师，文本分类，模型压缩

News Text Classification Method Based on Multi-Teacher Knowledge Distillation

Xiaojian Du*, Weidong Lv#, Yvhua Sun

School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou Gansu

Received: Jul. 5th, 2023; accepted: Aug. 3rd, 2023; published: Aug. 14th, 2023

Abstract

From traditional text classification to text classification based on deep learning, With the proposal of BERT model, it and its variants gradually become the mainstream model in natural language processing, but it needs to occupy and spend a lot of memory and computer resources. According to the dissimilarity of teacher-student network structure, it is dividing the two cases into isomorphic and heterogeneous teacher-student network, and proposes two different multi-teacher dis-

*第一作者。
#通讯作者。

文章引用: 杜潇鉴, 吕卫东, 孙钰华. 基于多教师知识蒸馏的新闻文本分类方法[J]. 计算机科学与应用, 2023, 13(8): 1515-1526. DOI: 10.12677/csa.2023.138150

tillation strategies. The experiment on the THUCNews dataset shows that even if there are teachers with poor performance, the classification effect of the student model can be improved by 3.26% and 3.30% respectively, and the performance loss is 0.79% and 0.78% respectively, indicating that the classification effect of the teacher model is close to that of the teacher model. At the same time, the number of participants is only 2.05% and 2.08% of the teacher model, which achieves a good model compression.

Keywords

Knowledge Distillation, Multiple Teachers, Text Classification, Model Compression

1. 引言

近年来，随着网络的快速发展，在丰富了广大人民生活的同时也使他们的生活得到了改变。人们更容易通过互联网获取信息，然而在互联网爆炸式发展的今天，网络上的文本有着数量庞大、内容复杂、类别繁多等特性，人们需要用大量的时间来过滤真正有意义的信息，从各式各样的文本数据中提取出需要的信息，渐渐成为人们日常生活的必然需求。

文本分类作为自然语言处理(NLP)中一个十分经典的应用，已经诞生几十年了，发展十分迅速，从传统的文本分类模型包括朴素贝叶斯、支持向量机、K 近邻、决策树、逻辑回归等，到深度学习下的文本分类包括 CNN 模型、RNN 模型、BERT 模型、Attention 机制的诞生，可以更好的利用词序的特征。随着样本和网络深度的增加，模型分类的精度越来越高，基于深度学习的方法已经在各项文本分类任务上，超越了传统的基于机器学习的方法，成为当前文本分类的主流方法。为了提高人们的用户体验，文本分类模型应该尽可能具有高的分类精度，同时文本分类模型的推理时间也必须满足响应速度快的要求。为了令分类模型在满足系统响应速度要求的前提下具有尽可能高的准确率，一种有效的方式是采用知识蒸馏(KD)来完成分类模型的训练，知识蒸馏的核心思想为：利用具有较强特征提取能力的大型深度学习模型对数据集中的深度知识进行学习，然后将这些深度知识转移到小型深度学习模型中，这样即使是小型深度学习模型也能获得接近教师网络的评价指标。

最近的一些变体尝试利用多个教师的知识来指导学生，鉴于多教师蒸馏多用来处理图像且给每个教师分配平均或者固定的权重[1] [2] [3]，而在文本分类方面的研究较少，本文提出基于交叉熵的多教师蒸馏的中文文本分类模型，并分成同构和异构两种，分别提出了不同的蒸馏策略。本文设置教师模型分别为预训练后的 BERT 模型和 BERT-CNN 模型，学生模型为 TextCNN 模型。实验结果表明，本文提出的方法比平均反映的多教师蒸馏的学生模型表现效果更好，同时比基于置信的多教师蒸馏(CA-MKD) [4]等一些方法效果好，比单独训练教师模型减少了参数量，说明其具有很好的压缩意义。

2. 相关研究

随着深度学习的发展，在自然语言处理(NLP)中更新换代了许多模型，对 NLP 任务的精度得到了提高。经典模型卷积神经网络(CNN)在深度学习中一般是用来解决计算机视觉和图像处理之类的问题，而 Yoon Kim 针对 CNN 做了一些变形，提出了一种更简单的模型，即 TextCNN 模型[5]。其中 TextCNN 模

型只有一层卷积层和最大池化层。在 TextCNN 之后又提出了可以解决时序问题循环神经网络(RNN) [6]，但由于 RNN 模型在梯度下降时容易出现梯度消失或梯度爆炸问题。于是针对梯度消失和梯度爆炸问题提出了长短期记忆网络(LSTM) [7]，LSTM 模型是在 RNN 的基础上增加了门控单元，在一定程度上提高了模型的性能。在 LSTM 提出后，Devlin 等人通过堆叠 Transformers 的编码部分提出了 BERT 模型[8]，引入了基于注意力机制[9]的多头注意力(Multi-Headed Attention)机制，同时考虑了当前词的前后单词对当前词的影响，从而能够生成更加准确的文本表征，进而提升文本分类准确率。同时，由于 BERT 模型能够并行地处理文本，使得其拥有更高的训练效率。

随着 NLP 任务在深度学习的发展和性能指标的改善，使 NLP 任务的产业化成为可能。然而昂贵的计算成本，使模型在移动设备上的部署变得困难。为了压缩模型以减少模型在计算上的时间和空间消耗。近年来模型压缩的主流方法有剪枝(Pruning) [10] [11] [12]、量化(Quantization) [13] [14] [15]、权重共享(Weight Sharing) [16]及知识蒸馏(Knowledge Distillation) [17] [18] [19]。近年来知识蒸馏备受关注，其在分类和预测任务中是非常有效的。Xinyin Ma 等人[20]提出了一种新的两级无数据蒸馏方法，用 12 层 BERT 模型作为教师网络用 8 层和 4 层 BERT 模型作为学生模型；廖胜兰等人[21]通过生成对抗网络得到的大量无标签数据，将教师模型的知识迁移到学生模型的一种知识蒸馏意图分类方法；Nityasya M.N.等人[22]用 BERT 模型或者混合 BERT 模型去指导 Bi-LSTM 模型和 CNN 模型在未标记数据集上实验。

不同于以上工作，本文认为多教师蒸馏也是具有研究意义的，提出了基于交叉熵的多教师知识蒸馏的中文文本分类模型，并分成同构和异构两种情况，对应的提出了不同的蒸馏策略，并取得了满意的实验结果。

3. 基础知识

3.1. 数据预处理

在进行模型训练之前，电脑需要把每个字都转换成向量。首先要对每句话进行分词操作，然后 BERT 模型和 TextCNN 模型有一份自己字典，其中样式为每个字前面都有一个索引，词嵌入层会把每个字用索引替代，然后从训练好的词向量中根据索引提取词向量，最终每个字都被转换为词向量。两种编码大致如下图 1 所示。

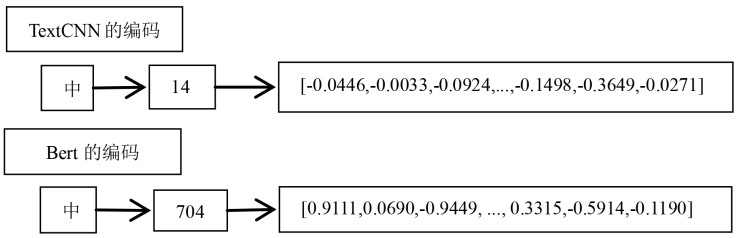

Figure 1. Partial word vector conversion of TextCNN and BERT
图 1. TextCNN 和 BERT 的部分词向量转换

其中 TextCNN 的词向量维度为 300 维，BERT 的词向量维度为 768 维。本文设置文本长度为 32，对于句子没达到长度时会默认填充[PAD]符号使得每句话有相同的长度。其中 BERT 编码的第一个代表[CLS]符号，预训练后的 BERT 模型通过微调生成词向量，而 TextCNN 模型本文选择用搜狗预训练词向量。

3.2. 教师模型

教师模型通常为复杂的、鲁棒性强的模型教师模型，简单来说就是一个大而深的高精度模型，本文

选用预训练后的 BERT 模型和 BERT-CNN 模型作为教师模型。

BERT 模型是把多个 Transformer 多个编码层堆叠起来，以便更好的特征提取。一般 BERT 模型有 12 层 Transformer 编码层，即 12 个块(block)。经过 BERT 编码过后可以作为词嵌入，后接入全连接层做分类器，如图 2 左边分叉所示。

BERT 在文本分类的应用：BERT 模型在每文本前会插入一个[CLS]符号，并将该符号对应的输出向量作为整篇文本的语义表示，用于文本分类，如下图 2 所示。也就是说[CLS]符号最后对应的输出包含了文本的全部信息，可以很好的做好分类任务。

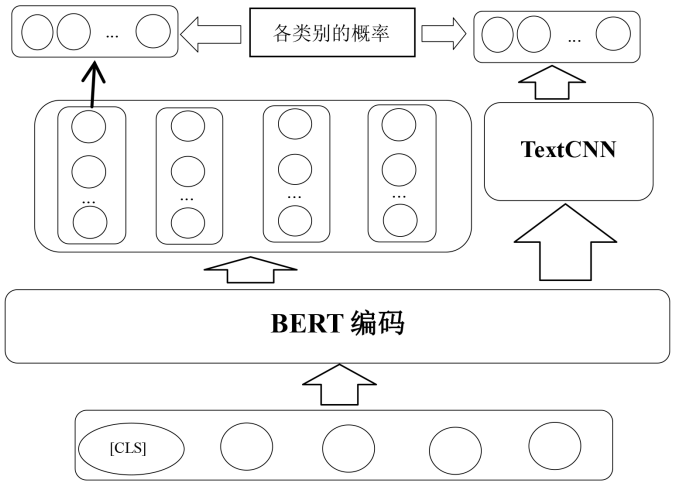

Figure 2. Network architecture of BERT and BERT-CNN models
图 2. BERT 和 BERT-CNN 模型的网络架构

其中右边分叉为 BERT-CNN 模型，即 BERT 作词嵌入后接入 TextCNN 模型做分类器。

3.3. 学生模型

教师模型会有好的表现，但模型一般都比较大，在训练过程中会用大量的内存和计算资源，就会使得模型运行对计算机内存和配置要求较高。这就需要一个简单快捷的网络来接受教师的知识。

本文选择 TextCNN 模型为学生模型。TextCNN 只有两层分别是卷积层和最大池化层。众所周知，卷积网络是在矩阵上进行滑动的，而 NLP 任务的输入是以矩阵表示的句子。与图像处理不同，卷积核的宽和每个词向量宽度一样，即每次读取都是一个字符，如下图 3 所示。

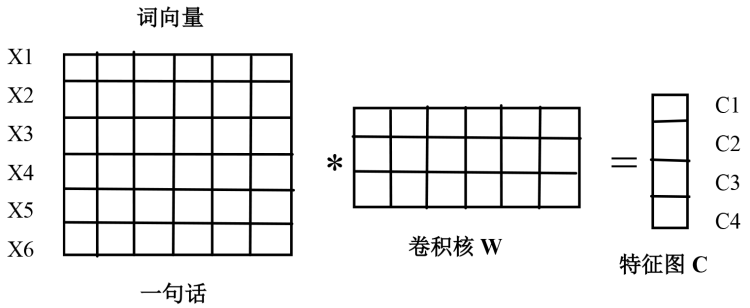

Figure 3. Calculation process of convolution
图 3. 卷积的计算过程

其计算公式如下，其中⊕是按行拼接，f为非线性激活函数：

$$X_{1:n}=x_1\oplus x_2\oplus\cdots\oplus x_n \tag{1}$$

$$c_i=f\left(W\cdot x_{i:i+h-1}+b\right) \tag{2}$$

$$c=\left[c_1,c_2,\cdots,c_{n-h+1}\right] \tag{3}$$

其中x是每个文字所对应的 TextCNN 的编码。卷积核的“宽度”就是输入矩阵的宽度(词向量维度)，“高度”可能会变，但一般是每次扫过 2~5 个单词，是整行进行的。一般来说，在大的词表中计算 3 元语言模型就会很吃力，所以本文按照 2，3，4 这样来取值。

4. 多教师知识蒸馏的文本分类

4.1. 多教师蒸馏模型

不同的教师体系结构可以为学生网络提供各自的有用知识。在培养学生网络的过程中，可以将多个教师网络单独或整体地用于蒸馏。在典型的师生框架中，教师通常有一个大模型或多个大模型的集合。要从多个教师那里转移知识，最简单的方法是使用所有教师的平均反应作为监督信号(Hinton *et al*., 2015) [17]。通常情况下，多教师知识蒸馏可以提供丰富的知识，并根据不同教师知识的不同，定制出一个全能的学生模型。

然而，如何有效整合多名教师的不同类型的知识。在日常学习中知道：同样的知识每个老师教学的结果是不同的，而权重是一个很好的参数，可以充分发挥每个参数的用处，本文认为在每个教师传递知识前乘上各自的权重，学生也会得到更好更全面的指导。

首先如何构造每个教师的权值，即每个老师对学生的指导力度，最期望的情况是教师越接近真值其指导效果越好，本文认为用教师和真值标签的交叉熵损失来判断各个教师的指导力度，教师与真值的损失公式为：

$$L_{T_{KD}}^{k}=-\sum_{i=1}^{N}y\ln\left(\sigma\left(z_{T_k}\right)\right) \tag{4}$$

$$w_{KD}^{k}=\frac{1}{K-1}\left(1-\frac{\exp\left(L_{T_{KD}}^{k}\right)}{\sum_{j}\exp\left(L_{T_{KD}}^{k}\right)}\right) \tag{5}$$

其中z_{T_k}是第k个教师网络的 Logits 输出，y是标签，k表示第k个教师，N表示数据中类别个数。

$$\sigma\left(z_i\right)=\frac{\exp\left(z_i/t\right)}{\sum_{j}\exp\left(z_j/t\right)}\ \left(t\text{表示温度}\right) \tag{6}$$

其中得到w_{KD}^{k}就是本文要的权重，可以看出在教师与真值的损失$L_{T_{KD}}^{k}$越小教师的指导力度w_{KD}^{k}就越大，反之教师表现($L_{T_{KD}}^{k}$)的越差指导力度(w_{KD}^{k})越小，则师生的蒸馏损失定义为：

$$L_{KD}=-\sum_{k=1}^{K}w_{KD}^{k}\sum_{i=1}^{N}z_{T_k}\ln\left(\sigma\left(z_S\right)\right) \tag{7}$$

其中z_S是学生网络的 Logits 输出，根据上面的公式可以看出教师表现的越好，其前面的权重就越大；反之权重就越低。相比于平均获取教师的知识，这样的构造学生就可以更好更全面的获取教师的知识。

教师的知识并不是什么都可以用来指导学生的，其中主要的知识有中间层特征知识传递和最后标签

知识传递，而在师生网络不相似时本文认为中间层特征的提取转换不如加强最后标签知识传递方便。所以本文考虑了两种情况下的多教师蒸馏分别为：同构多教师蒸馏和异构多教师蒸馏。

其中同构蒸馏是指教师和学生的模型架构相似或属于同一系列；异构蒸馏是指教师和学生的模型网络结构不完全相同、难以实现层间特征图匹配的情况。对于用 BERT 深度模型去指导 TextCNN 简单模型的本文认为是异构多教师蒸馏，即不选择用中间层特征做知识传递，选择加强最后标签知识传递；而对于用 BERT-CNN 深度模型去指导 TextCNN 简单模型的本文认为用同构多教师蒸馏，即选择用中间层特征和最后标签知识做知识传递。

4.1.1. 同构多教师蒸馏

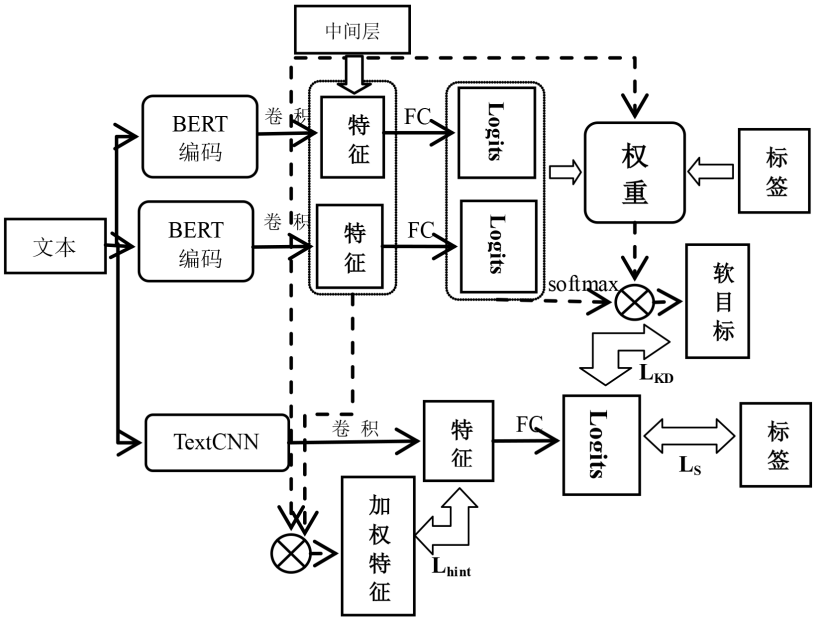

Figure 4. Isomorphic distillation network architecture
图 4. 同构蒸馏网络架构

如上图 4 所示是模型同构多教师蒸馏的网络构建，在确定师生蒸馏损失 L_{KD} 和学生损失 L_S 后，认为模型中间特征也是有指导意义的，考虑到 TextCNN 模型的简单性，本文采用教师和学生模型经过卷积池化后输出的特征作为中间层特征。在求得每个教师的权重之后，把其加入到教师对学生训练过程中进行中间层特征指导，本文选用光滑 L1 范数求得蒸馏误差：

$$L_{hint}=\begin{cases}\sum_{k=1}^{K} w_{KD}\left(f_{T_k}-f_S\right)^2 & \left|f_{T_k}-f_S\right|<1\\ \sum_{k=1}^{K} w_{KD}\left(\left|f_{T_k}-f_S\right|-0.5\right) & \text{otherwise}\end{cases} \tag{8}$$

其中 f_{T_k}, f_S 分别是第 k 个教师和学生网络的中间特征。光滑 L1 损失相比 L1 损失改进了零点不平滑问题。相比于 L2 损失，在 x 较大的时候不像 L2 对异常值敏感，是一个缓慢变化的损失。最后得出总的损失函数为：

$$L_S=-\sum_{i=1}^{N} y\ln\left(\sigma\left(z_S\right)\right) \tag{9}$$

$$L_{同} = L_S + \alpha L_{KD} + \beta L_{hint} \tag{10}$$

其中 L_S 是学生的训练损失；其中 α,β 是超参数，可以平衡蒸馏损失和中间特征损失以使得模型有更好的表现。

4.1.2. 异构多教师蒸馏

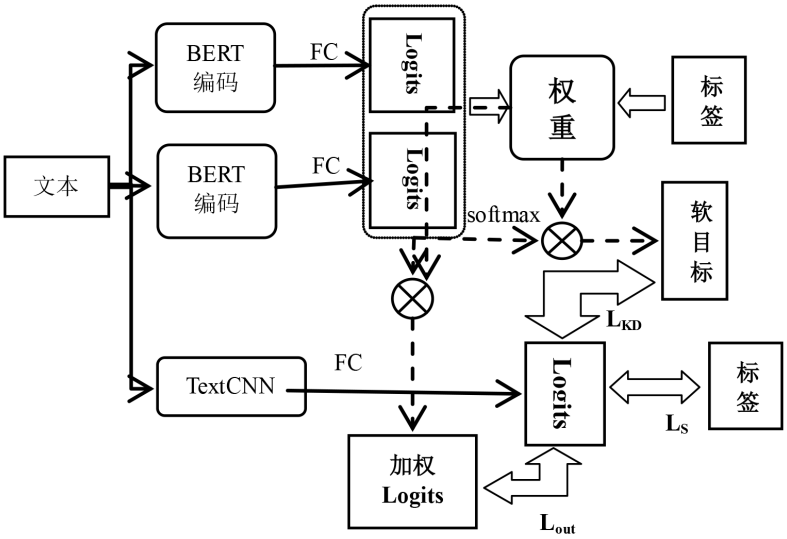

Figure 5. Heterogeneous distillation network architecture
图 5. 异构蒸馏网络架构

如上图 5 所示是模型异构多教师蒸馏时的网络构建，本文认为模型异构多教师蒸馏时，中间层特征的提取转换不如加强最后标签知识传递方便，所以本文考虑加强最后标签知识的指导意义。这里选择用 L2 损失函数来拉近学生和教师的 Logits 输出，同样也用之前得出的权重，最终得出以下公式：

$$L_{\text{out}} = \sum_{k=1}^{K} w_{KD}^{k} \left\| \sum_{i=1}^{N} z_{T_k} - z_S \right\|_2^2 \tag{11}$$

其中 z_{T_k} 和 z_S 分别是第 k 个教师网络和学生网络的 Logits 输出，最后得出总的损失函数为：

$$L_S = -\sum_{i=1}^{N} y \ln\left(\sigma\left(z_S\right)\right) \tag{12}$$

$$L_{异} = L_S + \alpha L_{KD} + \beta L_{\text{out}} \tag{13}$$

其中 α,β 是超参数，可以平衡蒸馏损失和标签知识损失以使得模型有更好的表现。

4.2. 实验与结果分析

4.2.1. 实验数据

在本节中，对 THUCNews 数据集进行实验，验证本文提出的多教师知识蒸馏的有效性。基于流行的神经网络架构，本文采用了预训练后的 BERT 模型、BERT-CNN 模型和 TextCNN 模型师生组合。

THUCNews 数据集在 14 个类别中抽取 10 个类别，每个类别抽取 10,000 条新闻标题，共 100,000 条，文本长度在 20 到 30 之间。类别：财经、房产、股票、教育、科技、社会、时政、体育、游戏、娱乐。分别定义为 0 到 9。数据集划分如下表 1 所示。

Table 1. Data set partitioning
表 1. 数据集划分

数据集	数据量
训练集	90,000
测试集	10,000

4.2.2. 评估方法

评估方法选取文本分类常用的指标：准确率(Acc)和 F1 得分作为模型效果的最终评价，公式为：

$$\mathrm{Acc}=\frac{1}{n}\sum_{i=1}^{n}\prod\left(y^{i}=\hat{z}^{i}\right) \tag{14}$$

$$F1=\frac{2\times P\times R}{P+R} \tag{15}$$

准确率表示预测正确的样本数占样本总数的比例，其中 n 为评估模型所使用的样本总数；$\prod(\cdot)$ 为指示函数，当输入的文本类别与预测的结果一致时，则指示函数的结果为 1，否则为 0；P 为查准率，R 为查全率。为了更好地体现模型压缩的有效性，在原有的评价体系上增加单位迭代次数所需要的时间(s)以及模型参数量两个评价维度。

4.2.3. 实验细节

本文的实验需要如下表 2 的配置。

Table 2. Experimental configuration
表 2. 实验配置

实验环境	实验配置
Python	3.8
PyTorch	1.10.2
CUDA	11.3+
GPU	Tesla P100

对于 BERT 和 BERT-CNN 模型均采用 BERTAdam 优化器，学习率设置为 0.00005，Warmup 设置为 0.05，中间隐藏层单元为 768，dropout 层为 0.1，epoch 设置为 3；在 TextCNN 模型采用随机梯度下降法，动量设置为 0.9，权值衰减为 0.0001，学习率设置为 0.001，卷积核分别用 2，3，4，卷积核数量为 256，dropout 层为 0.5。批大小均设置为 128。为了公平起见，在所有方法中，温度 T 设置为 5，而 α 设置为 3，epoch 设置为 5。在整个实验过程中，同构和异构分别将 β 设置为 10 和 1，随机数种子为 12。

4.2.4. 实验结果

通过在 THUCNews 数据集上实验，本文设计了一个表现好的教师和一个表现差的教师，以体现基于交叉熵的多教师蒸馏对比于平均接受每个教师的知识的好处，本文分别在同构和异构蒸馏的情况下做实验，得到如下表 3 和表 4 的结果。

Table 3. Experimental results of isomorphic distillation
表 3. 同构蒸馏实验结果

模型	Acc	F1 score
BERT-CNN (教师一)	94.07%	0.9123

Continued

BERT-CNN (教师二)	88.45%	0.8601
TextCNN	87.26%	0.8149
KD1 [17]	89.55%	0.8955
KD2 [17]	88.59%	0.8863
Avg Weight [1]	89.77%	0.8977
CA-MKD [4]	89.21%	0.8921
AEKD [23]	90.34%	0.9034
EBKD [24]	89.83%	0.8983
Our	**90.52%**	**0.9051**

其中 Avg Weight 是代表平均每个教师的指导力度，CA-MKD 是基于置信的多教师蒸馏，KD1 表示教师一进行单教师蒸馏，KD2 表示教师二进行单教师蒸馏，AEKD 表示基于梯度空间的自适应多教师蒸馏，EBKD 表示基于熵的多教师蒸馏。通过表 3 的实验结果可以看到，对于同构多教师蒸馏，Avg Weight 比学生模型 TextCNN 的精度提升了 2.51%；而 CA-MKD、AEKD 和 EBKD 精度分别提升了 1.95%、3.08%和 2.57%；但本文的方法对比与学生模型 TextCNN 精度提升了 3.26%，对比于 AEKD 和 CA-MKD 分别提升了 0.18%和 1.31%，说明本文采用交叉熵损失作为每个教师的指导力度和采用光滑 L1 范数来处理中间层知识传递的方法更优；并且可计算出本文方法蒸馏后相比于表现好的教师的性能损失(教师模型 − 蒸馏模型)/教师模型为 0.79%；通过与单教师蒸馏对比本文的方法在精度和性能损失方面也有所提高。

Table 4. Experimental results of isomerization distillation
表 4. 异构蒸馏实验结果

模型	Acc	F1 score
BERT (教师一)	94.09%	0.9128
BERT (教师二)	87.62%	0.8447
TextCNN	87.26%	0.8149
KD1 [17]	89.68%	0.8969
KD2 [17]	88.48%	0.8849
Avg Weight [1]	89.56%	0.8955
CA-MKD [4]	89.26%	0.8928
AEKD [23]	89.30%	0.8930
EBKD [24]	89.72%	0.8975
Our	**90.56%**	**0.9057**

而对于异构多教师蒸馏，从表 4 的结果可以看出，Avg Weight 比学生模型 TextCNN 的精度提升了 2.30%；同时 CA-MKD、AEKD 和 EBKD 使精度分别提升了 2.00%、2.04%和 2.46%；但本文的方法对比与学生模型 TextCNN 精度提升了 3.30%，比 EBKD 和 CA-MKD 分别提升了 0.84%和 1.30%，说明本文对于异构师生网络的蒸馏方法，即强化 Logits 输出做知识指导和提取中间特征的做知识指导效果更好，这样也避免了师生网络差别大时中间层的提取转换。同样也可计算出异构时本文方法蒸馏后相比于表现

好的教师的性能损失为0.78%，相对于单教师蒸馏本文在精度和性能损失方面也有所提高。

本文也对每个模型的参数量和训练后保存的模型大小进行了统计，得到如下表5结果。

Table 5. Comparison of the number of model parameters and the size of the training saved mode
表5. 模型参数量和训练保存模型的大小比较

模型	总参数量(百万)	保存模型
BERT-CNN	104.04	1.15G
BERT	102.27	1.14G
TextCNN	2.13	17.03M
Our	**2.13**	**8.52M**

从上表5可以看出相比于教师网络的BERT-CNN模型和BERT模型，学生TextCNN模型的总参数量和保存模型偏小。教师模型总参数量是学生模型的48.85和48.01倍，是训练出来模型的69.15和68.55倍。而本文的模型在总参数量上和学生模型一样的，同样只是教师模型的2.05%和2.08%，但训练出的模型却只是教师模型的0.72%和0.73%，即缩小了138.22和137.01倍，而且对比学生TextCNN模型压缩了2倍，认为本文的模型起到了很好的模型压缩效果。

同样本文也记录了模型的一次epoch时间，得到如下表6所示，可见本文的方法只用了原来21.05%的时间，同时相比于其他方法也更快。

Table 6. Comparison of model training once epoch time
表6. 模型训练一次epoch时间的比较

模型	一次epoch
BERT-CNN	1 h 35 min
AEKD [23]	50 min
EBKD [24]	21 min
Our	**20 min**

4.2.5. 消融实验

Table 7. Ablation experiments
表7. 消融实验

同构消融实验			异构消融实验		
模型	Acc	F1 score	模型	Acc	F1 score
Our (w/o L_{hint})	89.51%	0.8951	Our (w/o L_{out})	89.37%	0.8939
Our	**90.52%**	**0.9051**	**Our**	**90.56%**	**0.9057**

其中w/o表示没有。从上表7实验结果可知在师生同构时，w/o L_{hint}在没有提取中间层特征时，精度和F1得分会降低，说明中间层包含了蒸馏的有用信息。在师生异构时，w/o L_{out}在没有加强Logits输出特征的情况下，精度和F1得分也降低了，说明Logits输出和软目标一样包含了蒸馏的有用信息。

5. 结论

基于文本分类的问题，本文采用了多教师－学生的知识蒸馏模型，采用了预训练后的BERT模型和

BERT-CNN 模型作为教师模型，经典模型 TextCNN 作为学生模型，用以做多教师蒸馏。本文针对每个教师的指导力度采用了交叉熵的方式计算出的权重作为指导力度，同时把模型分成同构和异构两种情况并分别提出了用光滑 L1 范数来衡量师生中间层特征的距离和加强 Logits 输出的知识指导的蒸馏策略。在 THUCNews 数据集上进行实验，结果表明，本文在同构和异构的情况下即使部分教师表现较差，也使得学生模型分类效果分别提升 3.26%和 3.30%，且相比于表现好的教师性能损失分别为 0.79%和 0.78%，说明接近教师的分类表现，同时参数量只是教师模型的 2.05%和 2.08%；训练出的模型是教师模型的 0.72%和 0.73%，而且对比学生 TextCNN 模型也缩小了 2 倍，认为起到了很好的模型压缩效果。对于 THUCNews 数据集本文采用的是搜狗预训练词向量模型，这始终有所局限，下一步研究可以用更好的方法去提取文本中的特征作为词嵌入，以使模型精度再次得到提高。

基金项目

国家自然科学基金(11961039)。

参考文献

[1] You, S., Xu, C., Xu, C. and Tao, D.C. (2017) Learning from Multiple Teacher Networks. *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, 13-17 August 2017, 1285-1294. https://doi.org/10.1145/3097983.3098135

[2] Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J. and Ramabhadran, B. (2017) Efficient Knowledge Distillation from an Ensemble of Teachers. *Interspeech* 2017, 18*th Annual Conference of the International Speech Communication Association*, Stockholm, 20-24 August 2017, 3697-3701. https://doi.org/10.21437/Interspeech.2017-614

[3] Wu, M.-C., Chiu, C.-T. and Wu, K.-H. (2019) Multi-Teacher Knowledge Distillation for Compressed Video Action Recognition on Deep Neural Networks. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 2202-2206.

[4] Zhang, H., Chen, D. and Wang, C. (2022) Confidence-Aware Multi-Teacher Knowledge Distillation. *ICASSP* 2022-2022 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Singapore, 22-27 May 2022, 4498-4502. https://doi.org/10.1109/ICASSP43922.2022.9747534

[5] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, 25-29 October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[6] 杨丽, 吴雨茜, 王俊丽, 刘义理. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2): 1-6+26.

[7] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[8] Kenton, J. and Toutanova, L.K. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, Minneapolis, 2-7 June 2019, 4171-4186.

[9] Bahdanau, D., Cho, K. and Bengio, Y. (2015) Neural Machine Translation by Jointly Learning to Align and Translate. *The* 3*rd International Conference on Learning Representations*, San Diego, 7-9 May 2015, 1-15.

[10] Chin, T.-W., Ding, R.Z., Zhang, C. and Marculescu, D. (2020) Towards Efficient Model Compression via Learned Global Ranking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 1518-1528. https://doi.org/10.1109/CVPR42600.2020.00159

[11] He, Y.H., Zhang, X.Y. and Sun, J. (2017) Channel Pruning for Accelerating Very Deep Neural Networks. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1389-1397.

[12] Zhuang, Z.W., Tan, M.K., Zhuang, B.H., Liu, J., Guo, Y., Wu, Q.Y., Huang, J.Z. and Zhu, J.H. (2018) Discrimination-Aware Channel Pruning for Deep Neural Networks. *Annual Conference on Neural Information Processing Systems* 2018, *NeurIPS* 2018, Montréal, 3-8 December 2018, 875-886.

[13] Wang, K., Liu, Z.J., Lin, Y.J., Lin, J. and Han, S. (2019) Haq: Hardware-Aware Automated Quantization with Mixed Precision. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, *CVPR Workshops* 2020, Seattle, 14-19 June 2020, 8612-8620. https://doi.org/10.1109/CVPR.2019.00881

[14] Wu, J.X., Leng, C., Wang, Y.H., Hu, Q.H. and Cheng, J. (2016) Quantized Convolutional Neural Networks for Mobile Devices. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016,

4820-4828.

[15] Xie, Z., Wen, Z.Q., Liu, J., Liu, Z.Q., Wu, X.X. and Tan, M.K. (2020) Deep Transferring Quantization. 16*th European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 625-642. https://doi.org/10.1007/978-3-030-58598-3_37

[16] Pham, H., Guan, M.Y., Zoph, B., Le, Q.V. and Dean, J. (2018) Efficient Neural Architecture Search via Parameter Sharing. *Proceedings International Conference on Machine Learning*, Vol. 2, 4092-4101.

[17] Hinton, G., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. *Computerence*, **14**, 38-39.

[18] Romero, A., Ballas, N., *et al*. (2015) Fitnets: Hints for Thin Deep Nets.

[19] Yuan, L., Tay, F.E.H., Li, G.L., Wang, T. and Feng, J.S. (2020) Revisiting Knowledge Distillation via Label Smoothing Regularization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 3903-3911. https://doi.org/10.1109/CVPR42600.2020.00396

[20] Ma, X.Y., Shen, Y.L., *et al*. (2020) Adversarial Self-Supervised Data-Free Distillation for Text Classification.

[21] 廖胜兰, 吉建民, 俞畅, 陈小平. 基于BERT模型与知识蒸馏的意图分类方法[J]. 计算机工程, 2021, 47(5): 73-79.

[22] Nityasya, M.N., Wibowo, H.A., Chevi, R., Prasojo, R.E. and Aji, A.F. (2022) Which Student Is Best? A Comprehensive Knowledge Distillation Exam for Task-Specific BERT Models.

[23] Du, S.C., You, S., Li, X.J., *et al*. (2020) Agree to Disagree: Adaptive Ensemble Knowledge Distillation in Gradient Space. 34*th Conference on Neural Information Processing Systems* (*NeurIPS* 2020), Vancouver, 6-12 December 2020, 12345-12355.

[24] Kwon, K., Na, H., Lee, H., *et al*. (2020) Adaptive Knowledge Distillation based on Entropy. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 7409-7413. https://doi.org/10.1109/ICASSP40776.2020.9054698