基于BERT-CNN的新闻文本分类的知识蒸馏方法研究*

叶　榕，邵剑飞，张小为，邵建龙
（昆明理工大学 信息工程与自动化学院，云南 昆明 650500）

摘　要： 近年来，随着大数据时代进入人类的生活之后，人们的生活中出现很多无法识别的文本、语义等其他数据，这些数据的量十分庞大，语义也错综复杂，这使得分类任务更加困难。如何让计算机对这些信息进行准确的分类，已成为当前研究的重要任务。在此过程中，中文新闻文本分类成为这个领域的一个分支，这对国家舆论的控制、用户日常行为了解、用户未来言行的预判都有着至关重要的作用。针对新闻文本分类模型参数量多和训练时间过长的不足，在最大限度保留模型性能的情况下压缩训练时间，力求二者折中，故提出基于BERT-CNN的知识蒸馏。根据模型压缩的技术特点，将BERT作为教师模型，CNN作为学生模型，先将BERT进行预训练后再让学生模型泛化教师模型的能力。实验结果表明，在模型性能损失约2.09%的情况下，模型参数量压缩约为原来的1/82，且时间缩短约为原来的1/670。

关键词： 新闻文本；BERT；CNN；知识蒸馏
中图分类号： TP391.1　　**文献标志码：** A　　**DOI：** 10.16157/j.issn.0258-7998.223094

中文引用格式： 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.
英文引用格式： Ye Rong，Shao Jianfei，Zhang Xiaowei，et al. Knowledge distillation of news text classification based on BERT-CNN[J]. Application of Electronic Technique，2023，49(1)：8-13.

Knowledge distillation of news text classification based on BERT-CNN

Ye Rong，Shao Jianfei，Zhang Xiaowei，Shao Jianlong
(School of Information Engineering and Automation，Kunming University of Science and Technology，Kunming 650500，China)

Abstract： In recent years，after the era of big data has entered human life，many unrecognizable text，semantic and other data have appeared in people's lives，which are very large in volume and intricate in semantics，which makes the classification task more difficult. How to make computers classify this information accurately has become an important task of current research. In this process，Chinese news text classification has become a branch in this field，which has a crucial role in the control of national public opinion，the understanding of users' daily behavior，and the prediction of users' future speech and behavior. In view of the shortage of news text classification models with large number of parameters and long training time，the BERT-CNN based knowledge distillation is proposed to compress the training time while maximizing the model performance and striving for a compromise between the two. According to the technical characteristics of model compression，BERT is used as the teacher model and CNN is used as the student model，and BERT is pre-trained first before allowing the student model to generalize the capability of the teacher model. The experimental results show that the model parametric number compression is about 1/82 and the time reduction is about 1/670 with the model performance loss of about 2.09%.

Key words： news text；BERT；CNN；knowledge distillation

0　引言

随着大数据时代的到来，今日头条、新浪微博和豆瓣等主流新闻媒体APP产生海量新闻文本，因此如何将这些新闻文本进行快速有效的分类对于用户体验乃至国家网络舆情控制是十分必要的。针对中文新闻文本分类任务，研究者提出许多分类算法和训练模型，证明深度学习分类方法的有效性。

以BERT[1](Bidirectional Encoder Representation from Transformers)预训练模型为例：在文献[2]的实验中可以

* 基金项目：国家自然科学基金项目(61732005)

得出，BERT-CNN模型取得的效果最佳，但是从工程落地的角度来说，模型参数量过于巨大，仅仅一个BERT模型，参数就达一亿多。本文使用的是谷歌开源的面向中文的BERT预训练模型，占用内存大小为325 Mb。另一方面，针对训练时间过长的缺点，以该实验为例，训练18万条新闻文本数据消耗的时间为3.5 h，很显然对于未来的模型工程落地还存在很大的差距。因此，本文在保证不下降过多模型的准确率的前提下，将BERT-CNN进行模型压缩，降低模型体积以及模型的训练时间，提升模型的泛化能力。

本文创新点主要体现在：(1)对实验数据集进行了扩充处理，提升模型泛化能力；(2)通过观察不同的T和α的组合对模型蒸馏性能的影响确定最优组合值而不是固定值；(3)蒸馏场景不再局限于传统情感分析(二分类)，本实验面向10分类的文本分析，不同标签文本的蒸馏性能也不尽相同。

1 BERT的基础原理

BERT[1]是谷歌提出的一种基于深度学习的语言表示模型。当BERT被发布时，与11种不同的自然语言处理测试相比，取得最好的效果，它是NLP的重要研究成果。

BERT是一种基于语义理解的深度学习双向预训练的Transformer。BERT主要由5个核心部分组成：预训练、网络深度、双向网络、Transformer模型、语义理解。通过调研相关文献不难发现BERT是一个可以用作特征提取的双向预训练的深度学习模型。可以通过微调运用于下游任务，包括分类、回归、机器翻译、问答系统等。

2 CNN的基本原理

卷积神经网络(Convolutional Neural Networks, CNN)是一类包含卷积计算且具有深度结构的前馈神经网络(Feedforward Neural Networks)，是深度学习(Deep Learning)的代表算法之一。CNN主要由3个核心部分组成：输入层、隐含层、输出层。

(1)输入层：卷积神经网络的输入层可以处理多维数据，例如，一维卷积神经网络的输入层接收一维或二维数组。它与其他神经网络算法类似，使用的是梯度下降算法进行学习。

(2)隐含层：卷积神经网络的隐含层中包含卷积层、池化层和全连接层。其中，卷积层是对输入数据进行特征提；池化层是在卷积层进行特征提取之后，对输出的特征图进行特征选择和信息过滤。

(3)输出层和传统的前馈神经网络的输出层相同，在分类问题中它可以输出分类标签，在物体识别问题中它可以输出物体的中心坐标、大小和分类，在图像语义分割问题中它可以输出分类结果。

3 模型压缩

3.1 模型压缩的必要性

模型压缩就是在尽可能不改变模型效果的情况下减少模型的尺寸，使得模型有更快的推理速度。压缩后的模型与原始的模型类似，此外，在计算时只需要使用小部分的资源。下面从不同的角度来说明模型的效果。

无论是在新闻文本分类场景还是新闻文本情感分析场景，BERT-CNN模型取得的效果最佳，但是从工程落地的角度来说，模型参数量过于巨大，仅仅一个BERT模型，参数就达一亿多。另一方面，在一些实验中也发现训练时间过长的缺点，例如新闻文本实验[2]，训练18万条新闻文本(短文本，字符平均长度为20~30)数据消耗的时间为3.5 h，很显然对于未来的模型工程落地还存在很大的差距。因此，本文考虑将BERT-CNN进行模型压缩，在保证不下降过多模型的准确率的前提下，降低模型体积以及模型的训练时间，提升模型的泛化能力。

3.2 模型压缩的相关技术

在用于深度学习领域的模型压缩相关技术主要有4种：低秩分解、知识蒸馏、剪枝以及量化。

3.2.1 低秩分解

低秩分解(low-rank approximation)[3]主要目的是去除冗余和较少权值的参数，简单来说低秩分解是把原网络的连接权值矩阵当成满秩矩阵替换成若干个低秩矩阵，这几个低秩矩阵的组合逼近原始的连接权值矩阵，而每一个低秩矩阵又可分解成若干个较小矩阵的乘积，原先复杂而密集的连接权值矩阵也将被表示成较小规模简单矩阵的组合，从而实现结构简化的目的。

但是该方法也存在两个缺点：(1)低秩分解无法压缩一些卷积核本身就过小的网络；(2)模型被压缩后，模型的精度受损，需要重新训练。

3.2.2 剪枝

剪枝[4-6]就是在几乎不影响性能的情况下将已经训练好的神经网络模型里不重要的通道(神经元、连接权重和权重矩阵等)删除并对网络进行加速，剪枝主要有两种方式：(1)post-training剪枝[7]：模型无需再训练并在模型预测之前直接剪枝，但容易导致网络关键节点误删从而难以复原原模型的效果；(2)training剪枝[8]：在训练时剪枝，即使剪掉模型的重要内容也可以通过后续的模型训练恢复，但剪枝的自动化意味着更庞大的计算量。

3.2.3 量化

量化是将模型当中连续的权值进行离散化和稀疏化的过程。一般来说，神经网络建模的基本参数都是用宽度为32 bit的浮点型数来表达，但实际上实验结果并不需要保持那么高的精确度，因此就可以通过量化操作来降低模型的参数值。例如可以用0~255来代表32 bit所代表的精度，从而达到牺牲少量精确度来减少每个权

值占据的空间大小[9-11]。同时通过调研相关文献[9-11]，发现量化过程存在以下缺点：(1)操作复杂度大：在量化时需要做一些数值类型转换的处理，否则模型的精度损失会更严重；(2)通过微调的确能够减小精确度的损失，但是训练精度的确降低。

3.2.4 知识蒸馏

知识蒸馏的概念最早由Hinton等人[12]提出，该方法的核心思想就是先训练一个复杂的网络模型，然后通过训练好的模型参数值以及输出数据分布情况去训练一个更小更简单的网络。简单来说就是将大模型（教师模型）的学习结果作为小模型（学生模型）的训练目标，最后将教师模型的“能力”泛化到学生模型上，就此泛化后的学生模型代替教师模型进行测试。

3.2.5 模型压缩技术选择

通过对模型压缩的一些相关主流技术的介绍，将各个方法通过以浅显易懂且直观的方法进行对比，如图1所示。

图1　模型压缩技术对比

低秩分解：模型的轻量化主要通过替换权值矩阵的方法，但是当原模型存在一些本身就是低秩矩阵的卷积核便无法替换。

剪枝：通过拆剪一些神经元等不重要的通道实现压缩模型，但有可能会造成误删，使原模型的参数损失较大。

量化：实际操作需要大量的数值转换，操作难度大。

知识蒸馏：在保证原模型的性能不下降太多的同时将模型做轻量化处理，但是对数据的分布情况依赖性较大。

综上所述，考虑到计算复杂度、设备支持环境、模型压缩的应用场景以及现今各个模型压缩技术的优化情况，选择“知识蒸馏”作为模型压缩的方法。

4 知识蒸馏模型的原理以及构建方法

4.1 知识蒸馏的原理

假设在深度学习当中，从输入到输出存在一种未知的函数映射关系（黑匣子），那么，对于学生网络模型就需要通过原始数据集重新开始学习这个教师网络模型，也就是学生模型学习教师模型的泛化能力。一个例子：在之前[2]的新闻文本分类任务当中，当输入为“离谱！克罗德最后5秒超奇葩失误，葬送球队希望”模型训练的输出结果标签是“体育”，该结果表示样本在“体育”这个标签上有最大的概率值，同时，剩下的概率值则会分布到其他的标签（如“娱乐”“股票”等）。这些概率一般都很小，但是仍旧存在一些相关的信息，在这个文本样本中，“娱乐”的概率比“股票”概率大。因此，模型输出的标签识别信息更为丰富，信息熵也就越大，而这里的信息熵就是学生模型需要向教师模型学习的“经验”。

通过前面指出，“娱乐”和“股票”都有概率输出，但是“娱乐”的概率比“股票”概率大，不难看出不正确的标签的输出概率都非常小，这个概率趋近于0，因此这些不正确的标签输出概率对交叉熵损失函数的作用较低，在损失函数的作用中并没有被体现出来。为了让这些被忽略的信息被学生模型学习到，因此就需要以下两种方法[13-14]：(1)使用softmax（sigmoid函数）之前的值，最后计算损失函数；(2)将温度参数T加入到损失函数的计算，温度参数值越高，模型输出的概率曲线越平缓，因此可以得到“软标签”，从而进一步训练学生模型软化过程的公式：

$$q_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}} \tag{1}$$

其中，q表示学生模型需要学习的对象，也就是教师模型训练输出的软标签；z表示教师模型在经过softmax之前的逻辑回归值；T表示温度，T的作用就是放大小概率标签所包含的信息，当T=1时，表示所含的信息没有被放大；当T越高，教师模型的输出概率分布情况就越平缓，同时，小概率标签所包含的信息也就会被放大，更有利于学生模型学习到这些小概率标签包含的信息。

由于在训练过程中，教师模型虽然训练学习过的数据比学生模型多，但也有可能会“出错”，为了避免将这种错误让学生模型学习到，就需要适当加入一些“硬标签（真实数据的概率分布情况）”来降低错误被传递给学生模型的概率。因此，知识蒸馏的损失函数就是硬标签的损失函数与软标签的损失函数的结合[15]，如图2所示。

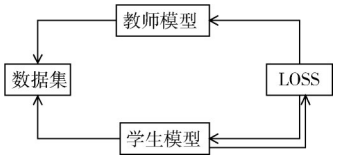

图2　蒸馏模型的损失函数结合过程

损失函数公式：

$$L = \alpha L^{(soft)} + (1-\alpha)L^{(hard)} \tag{2}$$

其中，α为权重，改变α时会对模型有不一样的影响，soft表示教师模型带着学生模型学习，然后将两者预测的结

《电子技术应用》http://www.chinaaet.com

果取交叉熵；hard表示学生看着“参考答案（真实数据标签）”学习，同样取二者的交叉熵。

4.2 知识蒸馏的训练过程

知识蒸馏总体的算法流程[16-18]如图3所示。

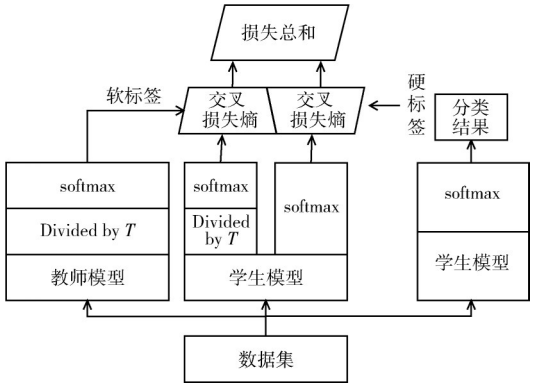

图3　知识蒸馏流程图

（1）通过一些带标签的数据训练教师模型；

（2）将训练好的教师模型计算出软标签；

（3）在同个T的条件下同样用带标签的原始数据训练学生模型，然后将输出的结果与步骤（2）得到的软标签进行交叉熵损失；

（4）令T=1，和步骤（2）得到的软标签进行交叉熵损失；

（5）将训练得到的学生模型进行原数据集预测，并与原始模型进行性能对比。

5 实验设计

5.1 实验模型设计

根据传统知识蒸馏的流程，将教师模型和学生模型分别替换为文中[2]训练好的BERT模型以及TEXT-CNN模型，总体蒸馏模型流程如图4所示。

基于以上的蒸馏模型，通过以下3种方法验证模型的有效性：（1）由于在蒸馏过程中，两个超参数：（T和α）是根据模型的表现情况进行确定的，因此需要观察不同的参数对模型性能的影响；（2）模型蒸馏后在同一个数据集的条件下与模型进行对比；（3）蒸馏后的模型在不同数据集的性能表现情况对比。

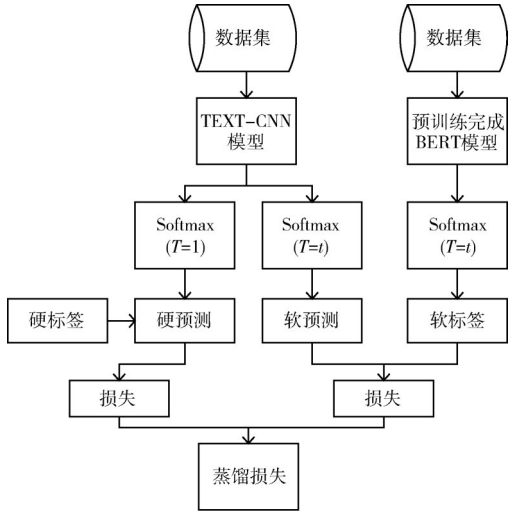

图4　BERT-CNN蒸馏流程图

5.2 实验数据集

为了让BERT-CNN模型具有更好的泛化能力，对原数据集部分进行数据扩充以及数据增强操作，具体操作如下：

（1）在原有的THUCNEWS数据集上，增加CLUE数据集，保留字符长度为20~30，总计25万条新闻文本数据，10个类别，每个类别2.5万条。将数据集按照8:1:1的比例拆分成训练集、验证集和测试集。

（2）采用BERT经典的mask方法，以一定的概率将[mask]随机替换文本当中的某一个单词。

（3）基于BERT的POS标签将文本中的单词以一定概率替换为随机采样的单词。

（4）随机从文本当中取出n-gram作为新的样本。

5.3 实验环境设置

采用计算机处理器为AMD-R5-3600六核十二线程，显卡为NVIDIA-GTX-1060（6 GB），基于Python-3.8，深度学习框架主要用的是Pytorch-1.9.0+cu102GPU版本，运行内存16 GB，由于显存容量限制，batch_size设置为16，Epoch设置为3。编程软件使用的是PyCharm社区版。

5.4 评价指标

为了能够更好地体现蒸馏后的模型的有效性和轻量化，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。

6 实验结果与分析

6.1 实验结果

（1）不同参数对模型性能的影响见表1。

通过表1可以看出：T=3，α=0.5时模型性能最佳，因

表1　各个参数组合对蒸馏模型的影响（F1-Score）

权重	T=1	T=2	T=3	T=4
α=0	0.752 2	0.760 1	0.762 2	0.761 3
α=0.1	0.753 5	0.762 1	0.768 7	0.764 3
α=0.2	0.755 6	0.765 6	0.770 1	0.772 1
α=0.3	0.757 5	0.767 9	0.779 8	0.775 4
α=0.4	0.759 3	0.769 2	0.780 4	0.778 8
α=0.5	0.763 7	0.780 2	0.798 1	0.785 5
α=0.6	0.769 6	0.779 5	0.797 3	0.784 1
α=0.7	0.771 9	0.775 4	0.783 3	0.775 4
α=0.8	0.772 1	0.772 3	0.774 5	0.769 3
α=0.9	0.795 5	0.761 3	0.762 3	0.758 1
α=1	0.815 1	0.762 2	0.754 4	0.752 4

此选择这组参数作为蒸馏时的固定参数。T=1，α=0：表示学生模型在没有教师模型的影响下训练真实数据集；T=1，α=1：表示原教师模型单独在没有学生模型的影响下训练真实数据集，也就是原模型BERT未蒸馏的效果。

同时，表1中实验结果表明：α从0~1呈现上升的趋势，以T=3为轴，两边的F1-Score出现下降的趋势，其主要原因是T值越大，说明小概率标签所包含的信息被放大得越大，学生模型对于这种错误信息的关注度就越大，因此才导致在蒸馏过程当中F1-Score下降，故通过参数调整实验可知，一方面确定蒸馏时模型最佳性能参数，另一方面表明T值并不是越大性能越好。

进一步提取表1的数据可知，学生模型、教师模型、蒸馏后的模型F1-Score的情况如表2所示。

表2　各模型实验结果（F1-Score）

模型名称	F1-Score
学生模型CNN	0.752 2
教师模型BERT	0.815 1
BERT-CNN蒸馏模型	0.798 1

通过表2可计算出BERT经过CNN蒸馏后的性能损失（教师模型-蒸馏模型）/教师模型为2.09%。

（2）模型在不同标签文本中的表现对比见表3。

表3　各模型在不同标签文本的实验结果（F1-Score）

不同标签文本	学生模型CNN	教师模型BERT	蒸馏模型BERT-CNN	性能损失/%
金融	0.723 2	0.817 4	0.789 8	3.376 6
房产	0.764 8	0.812 8	0.805 5	0.898 1
股票	0.771 9	0.813 5	0.795 1	2.261 8
教育	0.747 3	0.824 9	0.797 7	3.294 7
科技	0.755 9	0.811 7	0.808 9	0.344 9
社会	0.770 2	0.810 9	0.806 9	0.493 3
时政	0.737 8	0.821 2	0.799 4	2.654 7
体育	0.758 9	0.831 0	0.788 8	5.078 2
游戏	0.752 4	0.810 2	0.797 8	1.530 5
娱乐	0.739 6	0.797 4	0.791 1	0.790 1

表3实验结果数据表明：每个模型在各个标签的表现都不尽相同，学生模型CNN性能表现最佳为“股票”标签，最差为“金融”标签；教师模型BERT性能表现最佳为“体育”标签，最差为“娱乐”标签；蒸馏后的模型性能表现最佳为“科技”标签，最差为“体育”标签。其中，性能损失最大的为“体育”标签。

（3）以测试集2.5万条数据为例，蒸馏前以及蒸馏后完成一次迭代的时间对比见表4。

表4　蒸馏前后的迭代时间对比

模型类型	迭代一次所需时间/s
教师模型	623
蒸馏后的模型	0.93

6.2 结果分析

（1）模型大小：教师模型参数量为108.81百万，经过蒸馏后的模型参数量为1.32百万（将BERT蒸馏到CNN上，故计算CNN的参数量）。模型参数约为蒸馏前的1/82倍。

（2）训练时间：在同个测试集的条件下，教师模型迭代一次所需时间为623 s，经过蒸馏后的模型迭代一次所需时间为0.93 s，时间压缩为蒸馏前的1/670。

（3）F1-Score：总体上，蒸馏后的性能损失只有2.09%，性能损失较小且已经十分接近蒸馏前的性能。

值得注意的是，对于蒸馏后的模型，在不同标签的性能损失不同，性能损失最为严重的是“体育”标签，原因可能是被模型判定为“体育”的文本（例如“终于忍不住了！詹姆斯公开道歉！作出承诺”）包含其他标签（“社会”“娱乐”等）的文本，“詹姆斯”一词本身带有多种标签（“娱乐”“体育”等），在句子中蒸馏前其他标签的文本概率较小，而蒸馏后被放大，导致误判为其他标签。通俗而言，有些“体育”标签的词大概率会存在于“娱乐”标签的文本，但是“娱乐”标签的词出现在“体育”标签的文本概率就相对较小，这是一种社会现象，也是模型蒸馏后“体育”标签性能损失最严重的原因。

7 结论

本文针对实验中[2]性能表现最优的BERT-CNN模型存在的参数量巨大、迭代时间过长的情况提出进行模型压缩，而后通过各种模型压缩技术的优缺点对比，考虑到本文的实验对象、需要压缩的模型特点以及实验环境，选择知识蒸馏作为更符合本文的压缩技术。而后对数据进行扩充再对BERT进行微调，将BERT作为教师模型蒸馏到CNN模型当中，然后通过多个维度对蒸馏后的模型进行评估，最后实验结果表明在模型性能损失约为2.09%的情况下，模型参数量压缩约为原来的1/82且时间缩短约为原来的1/670，使改进的BERT-CNN模型进行工程应用落地成为可能。

参考文献

[1] 何凯.基于自然语言处理的文本分类研究与应用[D].南京：南京邮电大学，2020.

[2] 张小为，邵剑飞.基于改进的BERT-CNN模型的新闻文本分类研究[J].电视技术，2021，45(7):5.

[3] ZHOU G，CICHOCKI A，XIE S. Fast Nonnegative matrix/tensor factorization based on low-rank approximation[J]. IEEE Transactions on Signal Processing，2012，60

(6):2928-2940.

[4] ABADI M, AGARWAL A, BARHAM P, et al. TensorFlow: large-scale machine learning on heterogeneous systems[J]. arXiv:1603.04467v1, 2016.

[5] DE JORGE P, SANYAL A, BEHL H S, et al. Progressive skeletonization: Trimming more fat from a network at initialization[J]. arXiv:2006.09081, 2020.

[6] CARREIRA-PERPINÁN M A, IDELBAYEV Y. "learning-compression" algorithms for neural net pruning[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[7] LEE N, AJANTHAN T, GOULD S, et al. A signal propagation perspective for pruning neural networks at initialization[C]//Learning Representations, 2019.

[8] LEE N, AJANTHAN T, HS TORR P. Snip: single-shot network pruning based on connection sensitivity[C]//ICLR, 2019.

[9] BANNER R, NAHSHAN Y, HOFFER E, et al. Post-training 4-bit quantization of convolution networks for rapid-deployment[J]. https://arXiv.1810.05723,2018.

[10] CHOUKROUN Y, KRAVCHIK E, YANG F, et al. Low-bit quantization of neural networks for efficient inference[C]//IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019.

[11] FANG J, SHAFIEE A, ABDEL A H, et al. Post-training piecewise linear quantization for deep neural networks[J]. arxiv:2002.00104, 2020.

[12] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7):38-39.

[13] YIM J, JOO D, BAE J, et al. A gift from knowledge distillation: fast optimization, network minimization and transfer learning[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[14] TANG R, LU Y, LIU L, et al. Distilling task-specific knowledge from BERT into simple neural networks[J]. arXiv:1903.12136, 2019.

[15] HAIDAR MA, REZAGHOLIZADEH M. Text KD-GAN: text generation using knowledge distillation and generative adversarial networks[J]. arXiv:1905.01976, 2019.

[16] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: hints for thin deep nets[J]. arXiv:1412.6550,2014.

[17] CHENG Y, WANG D, ZHOU P, et al. A survey of model compression and acceleration for deep neural networks[J]. arXiv:1710.09282,2017.

[18] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer[J]. arXiv:1612.03928, 2016.

(收稿日期:2022-06-17)

作者简介:

叶榕(1998-),女,硕士,主要研究方向:自然语言处理。

邵剑飞(1970-),通信作者,男,硕士,副教授,主要研究方向:通信与信息系统,E-mail:1515346516@99.com。

张小为(1995-),男,硕士,主要研究方向:自然语言处理。

扫码下载电子文档

版权声明