计算机科学
COMPUTER SCIENCE

面向知识蒸馏的多助教动态设置方法

司悦航，成清，黄金才

引用本文

司悦航，成清，黄金才．面向知识蒸馏的多助教动态设置方法[J]．计算机科学，2025，52(5)：241-247.

SI Yuehang, CHENG Qing, HUANG Jincai. Multi-assistant Dynamic Setting Method for Knowledge Distillation [J]. Computer Science, 2025, 52(5): 241-247.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700059

面向知识蒸馏的多助教动态设置方法

司悦航[1]　成　清[1,2]　黄金才[1]
1 国防科技大学大数据与决策实验室　长沙 410073
2 湖南先进技术研究院　长沙 410072
(siyuehang@nudt.edu.cn)

摘　要　知识蒸馏在目标识别的模型压缩等关键领域受到重视。通过深入研究知识蒸馏的效率并分析教师模型和学生模型间知识传递的特点,发现合理设置助教模型可以显著缩小教师和学生之间的性能差距。然而,助教模型的规模和数量的不合理选择会对学生产生负面影响。因此,提出了一种创新的多助教知识蒸馏训练框架,通过动态调整助教的数量和规模,以优化知识从教师向学生传递的过程,从而提高学生模型的训练准确率。此外,还设计了一种动态停止知识蒸馏的策略,设置不同训练方法的学生模型作为对照组,实现对知识蒸馏停止回合的个性化设计,进一步提升学生模型的训练效率,并构建更精简高效的多助教知识蒸馏框架。通过在公开数据集上进行实验,证明了提出的面向知识蒸馏的多助教动态设置方法的有效性。

关键词:知识蒸馏;目标识别;多助教;动态设置;DSKD
中图分类号　TP301

Multi-assistant Dynamic Setting Method for Knowledge Distillation

SI Yuehang[1],CHENG Qing[1,2] and HUANG Jincai[1]
1 Laboratory for Big Data and Decision,National University of Defense Technology,Changsha 410073,China
2 Hunan Advanced Technology Research Institute,Changsha 410072,China

Abstract　Knowledge distillation is increasingly gaining attention in key areas such as model compression for object recognition. Through in-depth research into the efficiency of knowledge distillation and an analysis of the characteristics of knowledge transfer between the teacher and student models,it is found that the reasonable setting of an assistant model can significantly reduce the performance gap between the teacher and student. However,the unreasonable choice of the scale and number of assistant models can have a negative impact on the student. Therefore,this paper proposes an innovative multi-assistant knowledge distillation training framework,which optimizes the process of knowledge transfer from the teacher to the student by dynamically adjusting the number and scale of assistant models,thereby improving the training accuracy of the student model. In addition,this paper also designs a dynamic stopping strategy for knowledge distillation,sets student models with different training methods as a control group,and achieves personalized design of the stopping rounds for knowledge distillation,further improving the training efficiency of the student model and constructing a more streamlined and efficient multi-assistant knowledge distillation framework. Experiments on public datasets prove the effectiveness of the proposed multi-assistant dynamic setting method for knowledge distillation.

Keywords　Knowledge distillation,Object recognition,Multi assistants,Dynamic setting,DSKD

1　引言

在目标识别技术迅猛发展的当下,推理模型的能力实现了质的飞跃,推动了卷积网络模型规模的持续扩张。尽管如此,这些模型因计算复杂性极高,难以在分布式移动设备或嵌入式系统中得到广泛应用,这限制了其在众多下游应用场景中的实施。鉴于此,本文着眼于深度学习领域内的模型压缩问题,致力于探索利用规模较小的模型实现与原始模型相匹敌或接近的性能表现。模型压缩策略涵盖:设计高效率的神经网络架构以优化运算流程[1];对随机生成的大型神经网络进行权重修剪以降低其复杂度[2];执行数据剪枝以减少模型参数量[3];应用知识蒸馏技术,使小型网络能够复现大型网络的卓越性能[4]等。

知识蒸馏(Knowledge Distillation,KD)作为一种神经网络模型优化的关键技术,通过模型压缩促进了知识的高效迁移。该技术突破了传统基于类别标签的学习范式,创新性地实现了从参数丰富的教师网络到参数较少的学生网络的深层次知识传递。这种传递不仅包含了基本的标签信息,更涵盖了深层的内在特征和模式识别,极大地增加了学生模型的学习深度和广度。然而,教师网络与学生网络之间存在显著的

到稿日期:2024-07-09　返修日期:2024-11-07
通信作者:成清(chengqing@nudt.edu.cn)

规模或能力差异，这往往会导致学生网络在吸收和应用这些知识方面遇到问题。为应对这一难题，研究者们提出了一系列创新策略，包括引入助教模型以桥接知识传递的断层，或采用早期停止蒸馏技术以规避过度拟合的风险。这些策略的实施旨在提高知识蒸馏的效率与质量，确保学生网络在维持较小规模的同时，能够实现与教师网络相近的性能。

在深度学习领域，知识蒸馏的优化策略不断演进，其中基于助教(Teaching Assistant，TA)框架的模型压缩[5]已成为提升知识传递效率的关键途径。该框架通过在资深的教师网络与新兴的学生网络之间嵌入一个或多个中间层级的助教模型来弥合知识传递过程中的鸿沟。助教模型承担着从教师网络吸纳知识并转授给学生网络的职责，其角色犹如知识传递的纽带。然而，助教模型的过量引入可能引发信息过载或知识冲突等问题，对学生模型的学习效果产生负面影响。鉴于此，构建一个结构合理的助教框架显得尤为关键。本研究深入探讨了如何根据助教模型的不同规模和能力进行选择与配置，优化了教师、助教与学生三者之间的知识蒸馏策略，并实现了助教模型数量及其参与度的动态调节，从而为学生模型提供了更加丰富、易于吸收的知识资源，显著提升了学生的学习效率和训练成果。

早期，停止知识蒸馏(Early-Stop Knowledge Distillation)作为一种提升训练效率的改进策略，已被证明能够显著缩短训练周期，加快模型的收敛速度。本研究的发现进一步证实了这一点：在教师模型与学生模型规模相近的条件下，学生模型能够在较短的训练周期内迅速接近教师模型的准确率。相对地，在规模差异较大的环境下，学生模型则需要较长时间来吸收教师模型的知识内涵。这一现象揭示了模型结构相似性在促进知识快速传递中的重要性，为知识蒸馏过程中网络匹配度的优化提供了实践指导。

尽管早期停止知识蒸馏策略能够显著缩短模型训练周期，但如何精准确定其最佳停止时机，仍是深度学习模型压缩领域中一个亟待解决的问题。选择训练周期的 1/3 到 1/4 作为停止点，仅是一种粗略的策略设计，其有效性可能受限。而是否存在更为科学的设置方法，需要通过严谨的研究和大量的实验来加以验证。在本研究的实验基础上，发现助教数量的增加并不总是能直接提升模型的准确性或学习速度。助教性能的过剩，尤其是在数量过多时，可能会向学生模型传递误导性信息，妨碍其准确模拟教师模型的知识表示。因此，本文提出了一种面向助教模型数量和规模的优化方法，旨在提升学生模型的性能。此外，本文还设计了一种动态停止知识蒸馏策略(Dynamic Stop Knowledge Distillation，DSKD)，为知识蒸馏过程提供了重要的指导，以期达到更高效、更精确的模型训练效果。

本文的主要贡献归纳如下：

1)通过实验验证了助教模型规模对知识蒸馏效果具有显著影响，特别是发现助教模型规模与学生模型规模相近时，更有助于提高学生模型的准确性；

2)提出了一种多助教动态设置方法，其根据学生模型的训练进度，动态调整助教模型的数量和规模，以实现最优的知识传递效果，并通过多个数据集实验验证了其有效性；

3)设计了一种动态停止知识蒸馏策略(DSKD)，对多助教训练框架进行了改进，并深入探讨了学生模型的收敛回合与多助教蒸馏框架设计之间的相互关系，为模型压缩的实际应用提供支撑。

2 相关工作

知识蒸馏技术作为一项在模型优化领域内极具影响力的技术手段，已逐步成为小型学生模型模仿大型教师模型的关键策略。在神经网络的研究与应用实践中，知识蒸馏的概念被广泛采纳并深入探索。本研究基于 Hinton 等提出的框架，对知识蒸馏过程进行了系统性实施与评估。已有研究多聚焦于提升知识提取的效率与质量，并探索将这些技术适配至不同领域的可行性。例如，部分研究通过引入额外的损失函数[6]，致力于优化学生模型的内部特征图，以期更贴近教师模型的特征表现。同时，也有研究关注训练过程中的知识提取技术，或通过调整训练策略来培养性能更优的学生模型[7]。

知识蒸馏的适用领域持续扩展，其在文本分类[8]、序列建模[9]以及目标识别[10]等多个研究领域均展现出卓越的性能。然而，在模型训练的具体实践中，知识蒸馏同样面临着一系列挑战。特别是当教师与学生模型之间存在显著的能力差异时，学生模型可能难以完全吸收教师模型的知识精髓，进而影响训练的准确性与效率。为应对这些挑战，Zhang 等[11]提出了一种创新的方法以提升模型的精度，并显著减少了训练所需的时间。Kin 等[12]则开发了一种轻量级的随机森林算法，通过结合交叉熵与随机分组机制来缓解模型的过拟合问题。Heo 等[13]则认为，蒸馏过程中的约束不应仅限于神经元的激活值，而应扩展至激活区域。Wang 等[14]提出通过融合动态掩码注意力机制捕获来自多个教师模型的丰富特征，进一步丰富了知识蒸馏的策略与方法。

在知识蒸馏的领域内，提升模型性能的一种策略是实施重复知识蒸馏[15]，亦称为顺序知识蒸馏。该方法分阶段实施知识传递，旨在逐步优化模型的表现。研究表明，将网络在不同训练阶段的输出作为教师模型，可以有效地训练一系列网络，实现知识的有效传递。然而，本研究发现顺序知识蒸馏尽管在理论上具有显著优势，但在实际应用中，其效果并不总是优于传统的集成训练方法。原因在于，在连续的蒸馏过程中，网络可能会逐渐丢失一些关键的知识特征，导致顺序知识蒸馏在网络作为教师时的能力下降。此外，当学生模型的参数受限，或面对规模较小、质量有限的数据集时，学生模型可能难以充分吸收并有效利用从教师模型传递的知识，导致顺序知识蒸馏的实际效果不尽如人意。

另一种提升知识蒸馏效果的有效策略是引入助教模型。在学生网络与教师网络规模差异显著时，构建一系列中等规模的助教模型，可以有效地弥合两者之间的知识传递差距。Seyed 等[16]的研究发现，合理设计的助教模型能够显著提高学生网络的准确性，因为它们可以作为知识传递的中介。Hinton 在理论上也支持了这一观点，强调助教在缩小教师与学生能力差距中发挥的关键作用。这种方法能够高效地填补因规模差异导致的教师和学生之间的知识传递空白[17]。然而，关于如何确定助教模型的最佳规模和数量，仍需进一步的研究。

此外，Jang 等[18]提出了早期停止知识蒸馏的概念，这是一种旨在提高知识蒸馏准确率并节约训练资源和时间的策略。他们观察到，在知识蒸馏训练过程中，学生网络的收敛速度呈现出由快至慢的变化趋势。基于此发现，在学生网络训练的关键时刻停止知识蒸馏，转而采用其他训练方法，有助于减少时间和计算资源的消耗。他们在实验中验证了早期停止知识蒸馏的有效性，但采取了固定停止时机的方法，未能根据学生网络的实际训练状况进行动态调整，这限制了早期停止知识蒸馏策略的灵活性。为了克服这一局限性，对早期停止知识蒸馏策略进行了优化，根据学生网络的训练进度动态决定知识蒸馏的停止时机，以实现资源的最优分配和训练效率的最大化。

3 多助教动态设置的知识蒸馏方法

知识蒸馏是一种使简化模型通过模仿复杂模型来获得深层次数据理解的有效技术手段[19]。其核心优势在于能够利用教师网络经过充分训练后所积累的丰富信息。教师网络的输出不仅限于提供简单的标签信息，而是涵盖了对输入样本（例如图像）的全面分析与深入理解。在图像分类任务中，教师网络输出的概率分布不仅揭示了不同类别之间的细微差别，也反映了其对样本的全面认知。学生网络在模仿这些概率分布的过程中，能够学习到教师网络的推理逻辑和对数据的深层理解，而不仅仅是简单地复制最终的分类决策。这一过程促使学生网络超越了基本的标签映射，学习到了更为丰富和复杂的数据表示形式。

如图 1 所示，通过设置多个不同助教的方法，可以进一步优化知识蒸馏过程。图中右侧的曲线展示了不同的助教设置方案。这种多助教设置方法不仅丰富了知识传递的途径，也为学生网络提供了更为多样化的知识来源，并且提升了学生网络的精度。

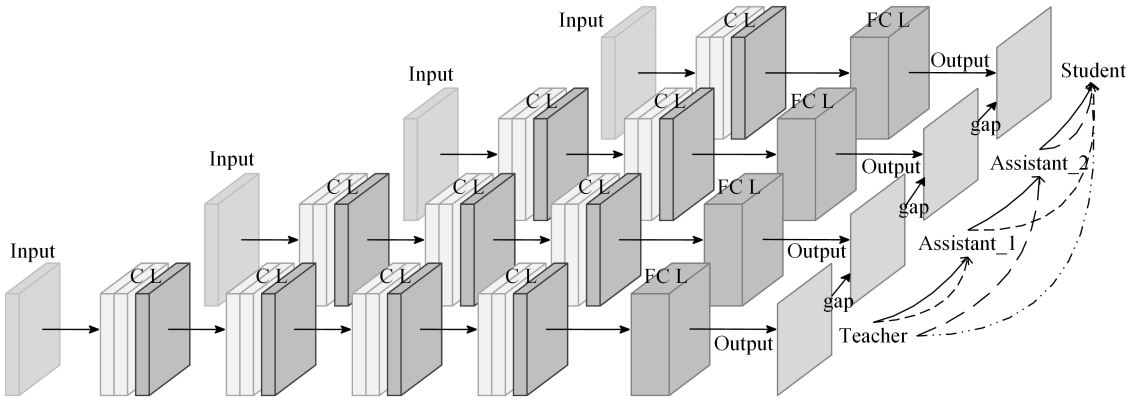


图 1 多助教模型的桥梁作用

Fig. 1 Bridge role of multi-assistant models

具体而言，将图像 x 作为输入，教师模型 k 产生得分的向量 $\boldsymbol{s}_k^t(x)=[s_1^t(x),s_2^t(x),\cdots,s_i^t(x)]$，$t$ 表示教师模型。训练后的神经网络产生一个峰值概率分布。知识蒸馏通过使用温度缩放来“软化”这些概率：

$$\widetilde{p}_k^t(x)=\frac{e^{\frac{s_k^t(x)}{h}}}{\sum_j e^{\frac{s_j^t(x)}{h}}} \tag{1}$$

其中，$h>1$ 表示温度，是稳定的超参数。

一个学生模型通过使用相同方法来创建软化的类概率分布 $\widetilde{p}^s(x)$。学生模型的损失是交叉熵损失 L_{cls} 和知识蒸馏损失 L_{KD} 的线性组合：

$$L=\alpha L_{\mathrm{cls}}+(1-\alpha)L_{\mathrm{KD}} \tag{2}$$

$$L_{\mathrm{KD}}=-\tau^2\sum_k \widetilde{p}_k^t(x)\log\widetilde{p}_k^t(x) \tag{3}$$

其中，α 和 τ 是两个超参数。一般情况下，$\tau\in\{3,4,5\}$，$\alpha=0.9$。

当使用知识蒸馏训练学生模型时，学生模型的准确率提高受限的一个至关重要的原因是，教师模型和学生模型之间的差距过大，学生模型无法完全学习到教师模型的知识。因此，在训练中设置助教可以有效地解决此问题。

教师模型和学生模型的性能差异计算如下：

$$R(f_s)-R(f_t)\leqslant O\left(\frac{\|F_s\|_C}{n^{\alpha_{st}}}\right)+\varepsilon_{st} \tag{4}$$

其中，$O(\cdot)$ 表示近似错误；$f_t\in F_t$ 表示教师模型；R 表示误差；$\|\cdot\|_C$ 是某个函数类的容量度量，这里是一个常数；n 是数据点的数量；$1/2\leqslant\alpha_{st}\leqslant 1$ 与学习率有关；ε_{st} 是教师模型 f_t 相对于学生模型 f_s 的近似误差。

本文在教师模型和学生模型之间设置了一个助教模型，并让助教向老师学习。通过式(4)可以得到：

$$R(f_s)-R(f_a)\leqslant O\left(\frac{\|F_s\|_C}{n^{\alpha_{sa}}}\right)+\varepsilon_{sa} \tag{5}$$

$$R(f_a)-R(f_t)\leqslant O\left(\frac{\|F_a\|_C}{n^{\alpha_{at}}}\right)+\varepsilon_{at} \tag{6}$$

其中，α_{at}，ε_{at}，α_{sa} 和 ε_{sa} 的定义与 α_{st}，ε_{st} 类似。因此，由 Seyed 等[16]的结论可知，设定助教的有效性可以通过式(7)证明。

$$\begin{aligned}&O\left(D_t+\frac{\|F_a\|_C}{n^{\alpha_{at}}}+\frac{\|F_s\|_C}{n^{\alpha_{sa}}}\right)+d_t+\varepsilon_{at}+\varepsilon_{sa}\\&\leqslant O\left(D_t+\frac{\|F_s\|_C}{n^{\alpha_{st}}}\right)+d_t+\varepsilon_{st}\\&\leqslant O(D_s)+d_s\end{aligned} \tag{7}$$

其中，D_t，D_s 表示教师和学生模型与真实目标函数之间的差异；d_t，d_s 表示对应的近似误差。通过这种方式，可以证明助教的存在可以提高学生模型的准确率。

4 动态停止知识蒸馏策略

在深入分析知识蒸馏的理论与实践后，本文提出了一种面向知识蒸馏的多助教动态设置方法，并设计了一种动态停

止知识蒸馏的训练策略 DSKD，旨在减少必要的训练回合，提高训练效率。

DSKD 策略的核心思想在于，在知识蒸馏的训练过程中，随着学生模型性能的逐步提升，其与教师模型之间的性能差距将逐渐缩小。这一变化将导致知识传递的效率逐渐降低，进而减缓训练进度。因此，及时地停止知识蒸馏不仅可以避免效率下降的问题，还能够根据学生模型的实际表现动态调整停止时机，这一动态设计具有重要的实践价值。基于 DSKD 策略，本文进一步设计了一种多助教蒸馏框架的动态设置算法，其伪代码如算法 1 所示。

算法 1 动态设置多助教蒸馏框架

```
Input: up: size of teacher; down: size of student
Output: AST: collection containing multiple assistants
1. size_AST[i]: size of i_th assistant in AST
2. N: Number of teaching assistants required by users
3. M. acc: accuracy of model M
4. K: A hyperparameter, control the stop epoch of KD
5. method=KD
6. EPOCHS=Epochs(a present parameter)
7. cmp_AST=new AST
8. k=0(a temporary parameter for counting)
9. for {i=1,i≤N}
10.   while {down≤size_ AST[i−1]≤size_AST[i]≤up}
11.     for {j=1,j≤EPOCHS}
12.       train(AST[i],method)
13.       if {cmp_AST[i]==available}
14.         train(cmp_AST[i],non_KD);
15.       end if
16.       if {AST[i]. acc ≤ cmp_AST[i]. acc}
17.       k=k+1
18.         if{k≥K}
19.           method=non_KD
20.             cmp_AST[i]=unavailable
21.         end if
22.       end if
23.     end for
24.     if {student[i]. acc≥student[i-1]. acc}
25.   set(AST[i])
26.     end if
27.   end while
28. end for
29. return AST
```

在助教模型的设置与优化过程中，DSKD 策略的应用不仅可以提高助教模型的准确度，还能够显著缩短其训练周期。为了精确判定 DSKD 策略的实施时机，设立对照组是至关重要的。这意味着在同一训练环境中，需要培养两个学生模型：一个接受基于 DSKD 策略的知识蒸馏训练，另一个则采用与教师模型相同的传统训练方法。通过对照组的设置，可以更准确地评估知识蒸馏的效果，并及时调整训练策略，以确保学生模型能够在最短的时间内达到最优性能。

在算法 1 中，第 12 行描述了一个关键过程：一个学生模型正通过知识蒸馏技术从教师模型中学习。相对地，在第 13 和 14 行，另一个学生模型并未采用知识蒸馏的方法，而是直接基于数据集中的样本进行学习。这两个学生模型在每个训练周期中同步进行训练，并对它们的准确率进行持续的比较和评估。在第 16 和 17 行中，当观察到未使用知识蒸馏训练方法的学生模型的准确率达到或逐渐超过采用知识蒸馏训练方法的学生模型时，记录次数。在第 18—20 行，当对比结果为“超过”，并达到一定次数时，应终止知识蒸馏的训练过程。这种及时的调整可以在不降低学生模型准确度的前提下，有效减少所需的训练周期，从而节约训练时间和计算资源。进一步地，在伪代码的第 24 和 25 行，当发现增加助教模型能够显著提升学生模型的性能表现时，当前的助教模型将被保留并继续用于训练。这一决策过程基于对学生模型性能提升的评估。

当一个助教模型构建并投入使用后，需要对助教模型效果进行评估。这一评估过程与学生模型的训练过程相似，目的是确定是否可以通过引入额外的助教模型来进一步提升学生模型的准确率。当新增助教模型无法提升性能时，表明多助教蒸馏框架的设置已经达到最优，此时应停止增加助教模型，以确保训练资源的高效利用和学生模型性能的最优化。通过这一系统性的方法，可以确保知识蒸馏过程中助教模型设置的科学高效。

5 实验分析

在本研究的实验阶段，选取了多种不同助教数量和规模的蒸馏框架，以深入探究助教模型对学生模型准确率的具体影响。同时，实验还研究了在不同训练阶段停止知识蒸馏过程对学生模型训练成效的作用。

5.1 实验设置

为了全面评估本文提出的多助教动态设计方法的性能，将其应用于 3 个广泛认可的公开图像分类数据集[20]：MNIST，CIFAR10 和 SVHN[21]。在 MNIST 数据集上，采用了全连接层模型（Multi Layer Perceptron，MLP），以检验该方法在简单图像分类任务中的有效性；在 CIFAR10 数据集上，采用了残差神经网络模型（Residual Neural Network，ResNet），以探索该方法在复杂图像识别中的性能；而在 SVHN 数据集上，则采用了卷积神经网络模型（Convolutional Neural Network，CNN），以评估该方法在处理具有挑战性的真实世界图像数据时的表现。所有参与实验的模型均从随机初始化的参数状态开始训练，以确保实验结果的客观性和可重复性。控制知识蒸馏的参数取值在 2～5 之间。在实验中，本文采用交叉熵损失函数作为评价模型性能的标准，该损失函数因在分类问题中的性能优越而被广泛采用。

5.2 与学生更相似的助教的性能更优

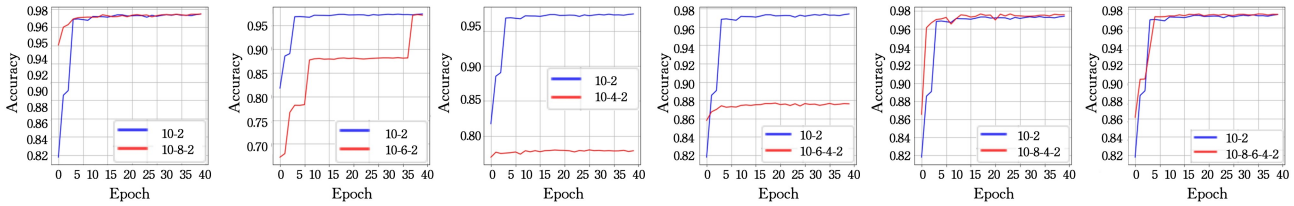
在前期研究的基础上，本研究创新性地提出了一种融合多助教动态设置与动态停止知识蒸馏（DSKD）策略的方法。为了实证该方法的有效性，在 3 个具有代表性的公开数据集上开展综合性能测试。为确保实验结果的可靠性与有效性，在实验设计中采用了与其他研究者相似的网络架构和参数配置，以最大限度地降低实验偶然性对结果的影响。实验结果的详细数据如表 1 所列。

表 1 知识蒸馏与本文方法的实验结果

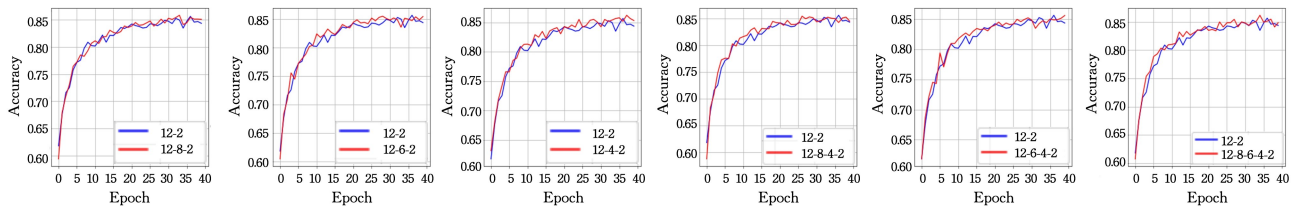
Table 1 Experimental results of knowledge distillation and the proposed method

数据集	模型	回合	方法	结构						
MNIST	MLP	30	—	10-2	10-8-2	10-6-2	10-4-2	10-8-4-2	10-6-4-2	10-8-6-4-2
			KD	97.43	96.69	97.21	97.41	97.33	97.26	97.36
			DSKD	97.62	97.44	97.31	78.02	87.68	97.68	97.68
CIFAR 10	ResNet	40	—	12-2	12-8-2	12-6-2	12-4-2	12-8-4-2	12-6-4-2	12-8-6-4-2
			KD	86.49	86.76	86.71	86.78	86.77	86.26	86.33
			DSKD	86.53	86.79	86.76	86.84	86.79	86.27	86.41
SVHN	CNN	40	—	12-4	12-10-4	12-8-4	12-6-4	12-10-6-4	12-8-6-4	12-10-8-6-4
			KD	92.67	92.09	92.99	92.92	90.56	91.44	92.04
			DSKD	94.70	93.84	94.45	94.44	94.07	94.35	94.63

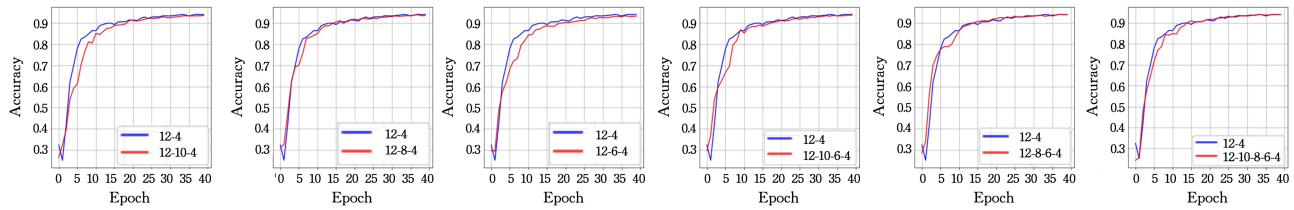
图 2 详尽地记录了不同助教框架在各个数据集上的测试过程和结果。在所有实验结果图中，深色线条代表了未使用助教模型的学生模型的训练轨迹，浅色线条则展示了采用所提方法的学生模型的训练进程。在训练的早期阶段引入知识蒸馏，当训练逐渐稳定后则停止知识蒸馏，直接在数据集上继续训练，以提升学生模型的学习效率。



(a)基于 MNIST 数据集的测试结果



(b)基于 CIFAR10 数据集的测试结果



(c)基于 SVHN 数据集的测试结果

图 2 知识蒸馏与本文方法的训练过程对比

Fig. 2 Comparison of training process between knowledge distillation and the proposed method

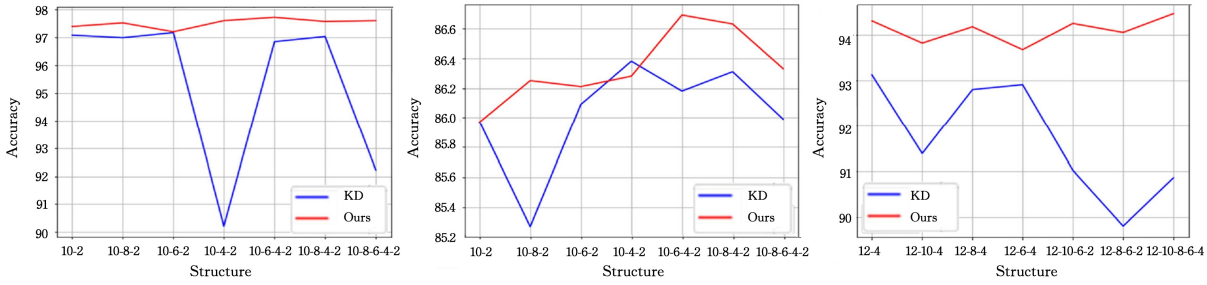
在 MNIST 数据集的实验中，经过 DSKD 策略训练的学生模型在平均准确率上比未经 DSKD 策略训练的模型约高出 3%。特别地，当多助教蒸馏框架按照“12-8-4-2”的规模设置时，学生模型达到了最高的准确率。对于 CIFAR10 数据集，经过 DSKD 策略训练的学生模型的准确率虽然有所提升，但提升幅度并不显著。这一现象主要归因于 CIFAR10 数据集的高复杂性以及 12 层教师网络的性能局限，这些因素共同作用，导致在知识蒸馏阶段未能充分发挥 DSKD 策略的优势。然而，当多助教蒸馏框架设置为“12-6-4-2”时，学生模型的准确率得到了最大程度的提升。在 SVHN 数据集上的实验中，接受 DSKD 策略训练的学生模型在平均准确率上比未接受 DSKD 策略训练的模型约高出 2.7%，且在多助教蒸馏框架设置为“12-10-8-6-4”时，学生模型取得了最高的准确率。

然而，在 MNIST 数据集上的 MLP 模型实验结果中，结构“10-4-2”和“10-6-4-2”出现了准确性远低于基线方法的现象，且在结构“10-6-2”中出现了准确性在训练后期才接近并超过基线方法的现象。造成这些现象的原因是，在实验中限制模型训练的最大次数为 30，这导致部分结构在训练后期才超过基线模型；部分结构甚至在预先设定的训练次数内都无法达到超过基线模型的效果。当不再限制模型训练的最大次数后，上述几个结构均能最终优于基线模型的效果。

5.3 DSKD 策略对知识蒸馏效率的提升

图 3 展示了 DSKD 策略在增强学生模型准确率和减少模型训练周期方面的显著效果。图中深色线条代表了在训练过程中应使用传统知识蒸馏方法的助教模型，而浅色线条则描绘了采纳早期停止知识蒸馏策略的助教模型。尽管在不同训练策略下，学生模型最终达到的准确率差异并不显著，但是当学生模型与教师模型在结构和参数量上相近时，学生模型往往能在较短的时间内实现显著的性能提升；相反，当两者

的参数量存在较大差异时，学生模型则需经过更长时间的训练才能达到与参数相近时相似的精度水平。



(a) NMIST (b) CIAFR10 (c) SVHN

图 3 KD 与 DSKD 对学生模型收敛回合的影响

Fig. 3 Influence of KD and DSKD on convergent epoch of student model

进一步地，通过对比 DSKD 策略的实施情况，可以观察到学生模型通常在训练的早期阶段能够更有效地吸收来自教师模型的知识。这表明在教师模型性能表现突出的初期，学生模型更易于获得快速的进步。然而，随着学生模型准确率的逐步提高，教师模型的优越性能逐渐减弱，难以持续有效地发挥其引导作用，这限制了学生模型进一步训练的潜力。因此，及时终止知识蒸馏的训练过程显得尤为关键，这不仅能够避免限制学生模型的学习潜力，还能促进其在不受教师模型现有知识限制的情况下，探索更深层次的知识，从而不断提升模型的精度。

此外，引入多个助教模型和设置双学生模型判断的机制会在一定程度上增加计算负担。为了解决这一问题，综合考虑模型复杂度与训练效率，设计 DSKD 策略有效减少训练回合，同时显著提升了学生模型的性能。在工业应用场景中，教师和助教模型的云端预训练及优化的模型传输策略，将进一步提高知识蒸馏效率并减轻用户端的计算压力。

5.4 多助教蒸馏框架对学生模型收敛回合的影响

图 4 给出了在 DSKD 策略影响下知识蒸馏训练持续回合数的直观展示。图中箭头上的数字具体指出了在 DSKD 策略影响下知识蒸馏阶段停止的确切回合数。观察结果表明，在学生模型与教师模型的参数规模相近时，学生模型往往能够在相对较少的回合内（约为总训练回合数的 25%）达到接近最终准确率的水平。这一现象突显了模型间结构相似性在促进知识快速传递和吸收中的关键作用。相反，当学生模型与教师模型的参数规模存在较大差异时，学生模型则需要更长的训练时间，约为总训练回合数的 50%，才能接近最终的准确率。

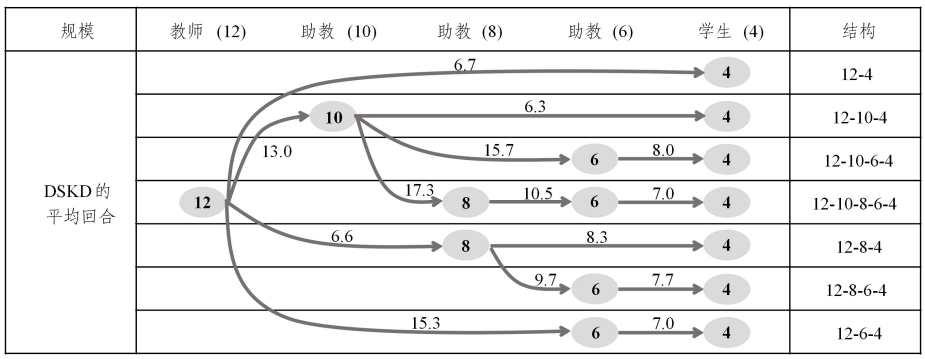


图 4 DSKD 策略在 SVHN 上的蒸馏回合路径

Fig. 4 DSKD distillation epoch path on SVHN

此外，尽管教师与学生模型的规模差异对 DSKD 策略所需的训练回合数有一定影响，但即便在这种情况下，所需的训练回合数依然显著少于传统知识蒸馏训练方法所需的回合数。训练回合数的显著减少，不仅展现了 DSKD 策略在知识传递效率上的优势，也彰显了其在加速学生模型训练进程中的重要性。

本文提出的多助教蒸馏框架方法在确保训练精度不受影响的前提下，大幅减少了模型的训练时间，显著提升了训练效率。该方法为深度学习模型压缩领域提供了一种新的优化途径，有助于在保证模型性能的同时，显著提高模型训练效率。所设计的 DSKD 策略和多助教蒸馏框架，为知识蒸馏技术的发展贡献了新的视角和实践方案。

结束语 本文研究发现，单纯增加助教模型的数量并非提升学生模型精度的普适解决方案。基于此，本文提出了一种创新的面向知识蒸馏的多助教动态设置方法及动态停止知识蒸馏（DSKD）策略。这些方法在模型参数层面有效地缩小了教师与学生之间的差距，并且显著提升了训练过程的效率。所提方法特别强调了助教模型群体间的协同以及与学生模型的知识传递。通过本文的设计，实现了模型间知识的高效传递，DSKD 策略确保了学生模型快速吸收并掌握关键知识。通过一系列实验证明了本文方法的有效性。

在未来的研究中，助教模型结构的多样化设计将为学生模型提供更为灵活的学习途径，特别是在面对特定领域知识时，能够更加便捷地进行训练。此外，知识蒸馏过程中损失

函数的优化调整，以及对学生模型共同训练产生的额外计算量的优化，也可能大幅提升知识传递的效率和质量。

参 考 文 献

[1] GORDON A，EBAN E，NACHUM O，et al. Morphnet：Fast & simple resource-constrained structure learning of deep networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：1586-1595.

[2] HAN S，MAO H，DALLY W J. Deep compression：Compressing deep neural networks with pruning，trained quantization and huffman coding[J]. arXiv：1510. 00149，2015.

[3] LIANG Z P，HUANG X J，LI S D，et al. Offline data-driven evolutionary optimization based on pruning stack generalization[J]. Acta Automatica Sinica，2023，49(6)：1306-1325.

[4] HINTON G，VINYALS O，DEAN J. Distilling the knowledge in a neural network[J]. arXiv：1503. 02531，2015.

[5] DONG X，HUANG O，THULASIRAMAN P，et al. Improved Knowledge Distillation via Teacher Assistants for Sentiment Analysis[C]// 2023 IEEE Symposium Series on Computational Intelligence(SSCI). IEEE，2023：300-305.

[6] ZAGORUYKO S，KOMODAKIS N. Paying more attention to attention：Improving the performance of convolutional neural networks via attention transfer[J]. arXiv：1612. 03928，2016.

[7] TARVAINEN A，VALPOLA H. Mean teachers are better role models：Weight-averaged consistency targets improve semi-supervised deep learning results[J]. arXiv：1703. 01780，2017.

[8] GUO W，HUANG J H，HOU C Y，et al. A text classification method combining noise suppression and double distillation [J]. Computer Science，2023，50(6)：251-260.

[9] FUKUDA T，KURATA G. Generalized knowledge distillation from an ensemble of specialized teachers leveraging unsupervised neural clustering[C]// ICASSP 2021－2021 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2021：6868-6872.

[10] SHI S H，WANG X D，YANG C X，et al. SAR image target recognition method based on cross-domain small sample learning [J]. Computer Science，2024，51(201)：465-471.

[11] ZHANG L F，SONG J B，GAO A，et al. Be your own teacher：Improve the performance of convolutional neural networks via self distillation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：3713-3722.

[12] KIM S，JEONG M，KO B C. Lightweight surrogate random forest support for model simplification and feature relevance[J]. Applied Intelligence，2022，52(1)：471-481.

[13] HEO B，LEE M，YUN S，et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019：3779-3787.

[14] WANG R Z，ZHANG X S，WANG M H. Text classification combining dynamic mask attention and multi-teacher multi-feature knowledge distillation[J]. Journal of Chinese Information Processing，2024，38(3)：113-129.

[15] YANG C L，XIE L X，QIAO S Y，et al. Knowledge distillation in generations：More tolerant teachers educate better students[J]. arXiv：1805. 05551，2018.

[16] MIRZADEH S I，FARAJTABAR M，LI A，et al. Improved knowledge distillation via teacher assistant[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：5191-5198.

[17] LIU S H，DU K，SHE C D，et al. Multi-teacher joint knowledge distillation based on CenterNet [J]. Systems Engineering and Electronics，2024，46(4)：1174-1184.

[18] CHO J H，HARIHARAN B. On the efficacy of knowledge distillation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：4794-4802.

[19] CHU Y C，GONG H，WANG X F，et al. Research on knowledge distillation algorithm for target detection based on YOLOv4 [J]. Computer Science，2022，49(201)：337-344.

[20] SHAO R R，LIU Y A，ZHANG W，et al. Review of knowledge distillation in deep learning [J]. Chinese Journal of Computers，2022，45(8)：1638-1673.

[21] GAO Y，CAO Y J，DUAN P S. Review on lightweight methods of neural network models [J]. Computer Science，2024，51(201)：23-33.

SI Yuehang，born in 2000，Ph. D. His main research interests include data fusion and knowledge processing.

CHENG Qing，born in 1986，associate professor，is a member of CCF (No. 31422G). His main research interests include knowledge reasoning and intelligence Q&A.

（责任编辑：何杨）