基于低秩自适应的伸缩感知蒸馏方法

李佳明[1],鲍志强[2],黄震华[1*],孙圣力[3],陈运文[4]
(1. 华南师范大学计算机学院,广东广州 510631;2. 恒生电子股份有限公司研究院,浙江杭州 310053;
3. 北京大学软件与微电子学院,北京 102600;4. 达观数据有限公司,上海 201203)

摘 要: 知识蒸馏是一种从复杂深层教师模型向轻量级学生模型迁移知识以提升性能的学习范式. 针对教师模型分布知识多样性不足,以及构建学生模型架构的搜索空间导致大量资源消耗的问题,本文提出了一种基于低秩自适应的伸缩感知蒸馏(Low-rank Adaptation based Flexibility-Aware distillation,LAFA)方法. LAFA方法通过构建低秩变换矩阵,将教师知识分别变换到学生模型的知识和类别标签,以提高分布知识的多样性. 同时,LAFA引入决策辅助器,动态伸缩学生模型容量,从而实现蒸馏性能与容量之间的均衡. 进一步,本文提出热启动和松弛策略来优化决策变量. 热启动策略通过约束学生模型缓慢增加容量,缓解因容量伸缩而导致的收敛困难. 松弛策略则在蒸馏后期移除约束,以少量资源消耗实现显著的性能提升. 在CIFAR-100数据集上,LAFA集成于13种蒸馏方法,平均性能提升了0.28个百分点. 同时,消融实验和分析实验进一步验证了LAFA方法的有效性.

关键词: 模型压缩;知识蒸馏;动态网络;模型正则化;深度学习
基金项目: 国家自然科学基金(No.62172166)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)04-1337-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240894

Low-Rank Adaptation Based Flexibility-Aware Distillation Method

LI Jia-ming[1], BAO Zhi-qiang[2], HUANG Zhen-hua[1*], SUN Sheng-li[3], CHEN Yun-wen[4]
(1. *School of Computer Science*, *South China Normal University*, *Guangzhou*, *Guangdong* 510631, *China*;
2. *Research Institute of Hundsun Technologies Co.*, *Ltd.*, *Hangzhou*, *Zhejiang* 310053, *China*;
3. *School of Software & Microelectronics*, *Peking University*, *Beijing* 102600, *China*;
4. *DataGrand Co.*, *Ltd.*, *Shanghai* 201203, *China*)

Abstract: Knowledge distillation is a learning paradigm that transfers knowledge from a complex and deep teacher model to a lightweight student model to enhance performance. To address the issues of insufficient diversity in the teacher model's knowledge distribution and the significant resource consumption caused by the search space for constructing the student model's architecture, we propose a low-rank adaptation based flexibility-aware distillation (LAFA) method. The LAFA method constructs low-rank transformation matrices to map teacher knowledge to both student model knowledge and class labels, thereby enhancing the diversity of distributed knowledge. Meanwhile, LAFA introduces a decision support module that dynamically adjusts the student model's capacity, achieving a balance between distillation performance and model capacity. Furthermore, we propose the warm-up and relaxation strategies to optimize decision variables. The warm-up strategy constrains the gradual increase in model capacity to alleviate convergence difficulties caused by capacity scaling, while the relaxation strategy removes the constraints in the later stages of distillation, achieving significant performance improvements with minimal resource consumption. On the CIFAR-100 dataset, LAFA integrated into 13 distillation methods achieved an average performance improvement of 0.28 percentage points. Moreover, through ablation experiments and analytical experiments, the effectiveness of the LAFA method is further validated.

Key words: model compression; knowledge distillation; dynamic network; model regularization; deep learning
Foundation Item(s): National Natural Science Foundation of China (No.62172166)

收稿日期:2024-10-07;修回日期:2025-01-24;责任编辑:覃怀银
*通讯作者:黄震华

1 引言

深度学习在计算机视觉、自然语言处理等人工智能领域已经取得巨大成功. 然而现有的深度学习模型往往被设计得深层而复杂,这限制了它们在轻量级设备上的应用. 另一方面,边缘智能的兴起也预示着人工智能逐渐与边缘设备深度融合,这要求模型更加的轻量且高效.

知识蒸馏[1]旨在利用深层复杂教师模型的知识来监督轻量级学生模型,提升学生模型在边缘设备上的性能. 目前,围绕知识类型和蒸馏流程,研究人员提出了众多方法. 在知识类型方面,文献[2]将教师模型的分布知识分解为两部分:目标类知识和非目标类知识,并引导学生分别学习这两部分知识. 文献[3]利用教师网络的浅层和深层特征共同指导学生网络的单层学习. 文献[4]通过为每个样本选取K个最近邻样本,构建局部相似性关系,并将样本间的邻域关系知识迁移到学生网络以提升其性能. 文献[1~4]假设教师模型的分布知识蕴含丰富的监督信息. 然而,上述方法存在监督知识多样性不足的问题,难以有效捕捉对学生模型学习最有利的知识,从而影响蒸馏性能. 在蒸馏流程方面,文献[5]提出利用一个“助教”模型来辅助蒸馏流程,缓解了师生模型之间规模差距过大、学生模型蒸馏学习困难的问题. 文献[6]利用NAS(Neural Architecture Search)技术构建学生模型架构的搜索空间,进而为当前教师模型匹配合适的学生架构. 但文献[5,6]需训练助教模型或大量中间模型,存在大量耗费资源问题.

为了解决上述问题,本文提出一种基于低秩自适应的伸缩感知蒸馏(Low-rank Adaptation based Flexibility Aware distillation,LAFA)方法,提高分布知识的多样性,同时通过伸缩学生模型的容量,实现蒸馏性能和容量的均衡. LAFA方法主要包含2个模块:低秩知识变换模块和伸缩感知模块. 低秩知识变换模块通过构建低秩变换矩阵,将教师知识分别变换到学生模型的知识和类别标签,以提高分布知识的多样性. 伸缩感知模块为学生模型配备决策辅助器,利用可微分的决策变量来约束学生模型的容量. 该模块根据当前师生模型之间的分布知识差距,对学生模型进行伸缩,从而实现蒸馏性能与模型容量之间的平衡. 本文中,基于现有工作[5],模型容量定义为模型的参数量.

此外,本文针对决策变量提出2种优化策略:热启动和松弛策略. 热启动策略约束学生模型在训练阶段缓慢增加容量,缓解其容量伸缩而难以收敛的问题. 松弛策略在蒸馏学习的后期移除对学生模型容量的约束,以少量资源消耗实现显著的性能提升. 为了验证所提出方法的有效性,本文在数据集CIFAR-100和ImageNet上进行了实验验证. 实验结果表明,LAFA方法作为一种即插即用的增强策略,能够集成到现有的知识蒸馏方法中显著提升蒸馏性能.

2 相关工作

2.1 知识蒸馏

文献[1]首次提出知识蒸馏概念,其主要思路是将预训练的模型作为教师模型,将其输出层的分布知识用于监督学生模型,提高后者性能. 因为教师模型结构和设计相对于学生模型通常较为复杂,具有很好的表示和泛化能力. 随后,研究人员进一步探索知识的类型以及蒸馏流程,并被广泛应用在各种领域.

在知识类型方面[1~4,7,8],其基本思路为建模教师模型在训练或推理阶段中的“知识”并指导学生进行蒸馏学习. 文献[7]提取教师模型的中间层特征作为蒸馏知识,并约束这些特征经过网络变换维度之后对齐学生模型的中间层特征,进而提高学生模型性能. 文献[8]挖掘师生模型中间层特征的互信息,以此促进学生模型学习鲁棒的特征表示. 在蒸馏流程方面[5,6,9],其基本思路为在预训练教师模型指导学生蒸馏学习的基础上,发展更有效的蒸馏流程. 文献[9]采用投票策略,从多个教师网络中筛选出相对差异化的中间特征知识,以提升学生模型性能.

此外,知识蒸馏被广泛应用在各种不同领域[10,11]. 文献[10]构建了变分贝叶斯编码器交通流预测模型,通过教师模型指导学生模型学习,显著提升了模型的预测性能和鲁棒性. 文献[11]提出了自蒸馏HRNet目标分割方法,通过结构化自蒸馏学习模型,并引入多尺度池化金字塔表示模块,有效提升了网络的目标分割性能,且无需增加资源开销.

2.2 动态模型

大多数深度学习模型以静态方式进行训练和推理,这在很大程度上限制了其计算效率. 基于此,研究人员考虑如何在不牺牲模型性能的前提下,控制模型的中间运算过程来降低模型容量. 文献[12]提出SkipNet,利用门控组件的可微分随机决策变量来决定是否跳跃卷积层. 而在知识蒸馏中固定的教师模型设定下,研究人员提出搜索合适的学生模型来适配教师. 基于NAS的方法[6]构建全局搜索空间来获取合适的学生架构,导致大量资源消耗. 受到动态模型的启发,本文提出了一种插入决策辅助器的学生模型,以适配教师模型,无需构建搜索空间.

3 LAFA方法

3.1 预备知识

假定存在2个神经网络模型:一个是具有大规模参

数且性能优异的教师模型;另一个是具有小规模参数且性能较差的学生模型. 在学生模型输出分布 q 与类别标签 y 之间的交叉熵损失 $\mathcal{L}_{\mathrm{CE}}$ 之外,知识蒸馏的目标为最小化教师模型输出分布和学生模型输出分布之间的 KL(Kullback-Leibler)散度 $\mathcal{L}_{\mathrm{KL}}$,其蒸馏总目标函数为

$$L=\mathcal{L}_{\mathrm{CE}}+\mathcal{L}_{\mathrm{KL}}(p,q) \tag{1}$$

其中,p 为教师模型输出,在蒸馏学习中被视为一种除类别标签之外的监督知识.

3.2　方法概述

LAFA 方法的框架如图 1 所示,样本作为输入分别送入预训练的教师模型和随机初始化的学生模型中进行蒸馏学习. 低秩知识变换模块构建高效的低秩变换矩阵,将教师模型的输出分布分别变换到学生模型的输出分布以及类别标签. 伸缩感知模块在学生模型中插入了决策辅助器,其中可微分的决策变量根据当前师生模型之间的分布差距,动态伸缩学生模型的容量.

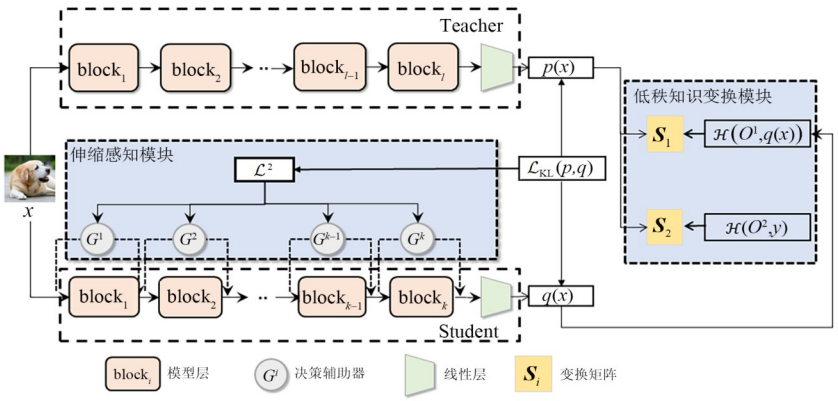


图 1　LAFA 方法总体架构图

3.3　低秩知识变换模块

为了提高教师模型分布知识的多样性,进而捕捉有利于学生模型学习的监督知识,本文提出了低秩知识变换模块. 该模块通过随机初始化一个参数矩阵,将教师模型的分布知识 p 映射到学生模型的输出分布 q. 通过训练该矩阵,拟合教师和学生模型之间的分布关系. 训练后的矩阵生成的分布知识相较教师模型分布知识 p 更为平滑. 教师模型的分布知识通常具有较高的确定性,这使得学生模型能够明确地学习类别信息. 但也可能导致学生模型对教师模型预测过于自信,从而容易发生过拟合. 通过参数矩阵变换得到的平滑分布,能够惩罚教师模型过于自信的输出,减少过拟合并增强学生模型的鲁棒性. 具体来说,对于样本 x,教师和学生模型的输出分布分别为 $p(x)\in\mathbb{R}^{1\times K}$ 和 $q(x)\in\mathbb{R}^{1\times K}$,$K$ 为类别数目. $p(x)$ 来自

$$p_i(x)=\frac{\exp\left(\frac{z_i}{\tau}\right)}{\sum\exp\left(\frac{z_i}{\tau}\right)} \tag{2}$$

其中,z_i 为模型逻辑层的第 i 个输出,$p_i(x)$ 为其经过 softmax 函数后的概率值. 温度参数 τ 用于控制分布的平滑程度. 在得到教师和学生模型对应的输出分布之后,引入针对 $p(x)$ 的变换矩阵 $\boldsymbol{S}_1\in\mathbb{R}^{K\times K}$,以 $q(x)$ 作为变换目标计算 KL 散度,对 $\boldsymbol{S}_1$ 的损失项为

$$\mathcal{H}(O^1,q(x))=\sum q_i(x)\lg\left(\frac{q_i(x)}{O_i^1}\right) \tag{3}$$

其中,O_i^1 为教师模型的第 i 个输出经过 $\boldsymbol{S}_1$ 变换后的对应输出. 与传统方法中将教师模型的知识蒸馏给学生模型的方向不同,变换矩阵 $\boldsymbol{S}_1$ 以学生模型输出分布为训练目标,这一方向转换使得该矩阵能够捕捉有利于当前学生模型蒸馏学习的知识. 此外,变换矩阵 $\boldsymbol{S}_1$ 可被视为一种优化矩阵,控制教师模型输出的概率分布.

此外,本文构建以 $p(x)$ 为输入,以类别标签为约束目标的变换矩阵 $\boldsymbol{S}_2\in\mathbb{R}^{K\times K}$. 相对于传统蒸馏中利用单一的教师输出分布作为监督知识,$\boldsymbol{S}_2$ 利用变换能力构造出额外的监督知识,提高了知识的多样性. 对应损失项为

$$\mathcal{H}(O^2,y)=\sum y_i\lg\left(\frac{y_i}{O_i^2}\right) \tag{4}$$

其中,O_i^2 为教师模型的第 i 个输出经过 $\boldsymbol{S}_2$ 变换之后的对应输出;y_i 为类别标签在第 i 个类的标签. 该损失项优化 $\boldsymbol{S}_2$,可被视为对 $p(x)$ 的变换操作,其中 $\boldsymbol{S}_2$ 的每一行都自适应调整 $p(x)$ 对每个类别的置信度,使得其更加接近于独热编码形式.

然而,$\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 矩阵面对高维输入输出会出现维度爆炸问题. 本文针对此问题对 $\boldsymbol{S}_1$ 和 $\boldsymbol{S}_2$ 施加低秩约

束[13],降低其维度计算. 假设$\boldsymbol{S}$矩阵具有低秩结构,即存在$\boldsymbol{L},\boldsymbol{L}_r\in\mathbb{R}^{K\times r}$,满足$r\ll K$且$S=\boldsymbol{L}\boldsymbol{L}_r^{\mathrm{T}}$. 利用2个低维子矩阵$\boldsymbol{L}$和$\boldsymbol{L}_r^{\mathrm{T}}$避免了直接处理高维矩阵$\boldsymbol{S}$本身. 当$r$趋向于1时,$\boldsymbol{L}\in\mathbb{R}^{K\times 1}$,这表示$\boldsymbol{L}\boldsymbol{L}_r^{\mathrm{T}}$当前对教师模型输出分布的变换能力较弱,其计算复杂度为$\sigma(K)$. 当r趋向于K时,$\boldsymbol{L}\in\mathbb{R}^{K\times K}$,其对教师模型输出分布的变换能力较强,但其计算复杂度为$\sigma(K^2)$. 为了让上述变换$\boldsymbol{S}_1$和$\boldsymbol{S}_2$能够捕捉教师模型输出分布中的隐式结构信息,本文对所有的低秩矩阵$\boldsymbol{L}$添加了随机dropout层.

$$\tilde{\boldsymbol{L}}=\mathrm{dropout}(\boldsymbol{L};\ v) \tag{5}$$

其中,v表示dropout的参数. 综上所述,低秩知识变换模块的损失函数为

$$\mathcal{L}^1=\alpha\sum q_i(x)\lg\left(\frac{q_i(x)}{O_i^1}\right)+\beta\sum y_i\lg\left(\frac{y_i}{O_i^2}\right) \tag{6}$$

其中,α和β是控制$\boldsymbol{S}_1$和$\boldsymbol{S}_2$远离固定的教师模型输出分布的权重超参数,调整监督知识的多样性程度.

3.4 伸缩感知模块

为了实现学生模型架构的动态伸缩,从而在蒸馏性能和学生模型容量之间达到平衡,本文提出了伸缩感知模块. 该模块在学生模型中插入决策辅助器,根据师生模型分布知识差距进而约束学生模型容量. 具体来说,首先在学生模型所有的可跳跃层配备可微的离散决策辅助器,通过损失函数优化其中的决策变量来决定是否跳跃该层的运算. 可跳跃层表明该层网络的输入可直接跳跃该层的计算过程,作为下一层的输入. 为实现该层,定义$f^i(x^i)$为学生模型第i层对其输入x^i的运算,则第$i+1$层的输入为

$$x^{i+1}=G^i\left(x^i\right)f^i\left(x^i\right)+\left(1-G^i\left(x^i\right)\right)x^i \tag{7}$$

其中,$G^i(\cdot)\in\{0,1\}$为在第i层的辅助器函数. 其中决策策略定义为

$$\Pi\left(x^i,i\right)=F\left(x^i\right) \tag{8}$$

$$g=\left[g_1,\cdots,g_l\right]\sim\Pi_{f_\theta} \tag{9}$$

其中,f_θ为模型参数的序列,l为模型的层数. 该决策策略首先将x^i送入到线性层$F(\cdot)$中,再基于输出$F\left(x^i\right)$构建概率空间$\Pi\left(x^i,i\right)$. 本文利用Gumbel-max采样方法[14]从$\Pi\left(x^i,i\right)$中采样决策变量g_i. 接着,经过指令函数运算后决策变量输出0或1,将其与输入x^i相乘,完成是否跳跃该层的决策. 具体的跳跃决策取决于学生模型前向传播中的硬门控机制和反向传播中的软门控机制:

$$G^i\left(x^i\right)=\begin{cases}I\left(g_i\geqslant 0.5\right),\text{forward}\\ g_i,\text{backward}\end{cases} \tag{10}$$

其中,$I(\cdot)$为指令函数. 举例来说,在第i层的前向传播运算中,决策辅助器从$\Pi\left(x^i,i\right)$采样出g_i,当$I\left(g_i\geqslant 0.5\right)$时,$x^i$不跳跃该层运算,反之跳跃. 在第$i$层的反向传播运算中,$g_i$被损失函数正常优化. 对$g$施加的约束如下:

$$\mathcal{L}_g=\sum_{i=1}^{l}\left(-\left(1-g_i\right)\right) \tag{11}$$

配备决策辅助器的学生模型相比于传统的静态学生模型,可根据输入数据的差异动态调整模型容量. 值得注意的是,可跳跃层的限制之处在于其要求该层的输入维度等于输出维度. 对于某些要求维度变换的层并不适用跳跃操作.

在此基础上,根据蒸馏动态来优化g_i. 具体地,引入师生模型之间分布差距来改进式(11):

$$\mathcal{L}^2=-\gamma\left[\left(\frac{1}{\sum q_i(x)\lg\left(\frac{q_i(x)}{p_i(x)}\right)}\right)\sum_{i=1}^{l}\left(-\left(1-g_i\right)\right)\right] \tag{12}$$

其中,γ为超参数,$\sum q_i(x)\lg\left(\frac{q_i(x)}{p_i(x)}\right)$表示当前训练过程中师生模型输出分布之间的损失值,本文将其倒数建模为蒸馏动态. 学生模型在损失值较大时,减低对g_i的约束,进而更好地蒸馏教师的分布知识.

此外,为优化决策变量,提出2种优化策略:热启动和松弛策略.

(1)首先引入t比例系数来约束$\frac{1}{l}\sum_{i=1}^{l}g_i$. t越大,约束跳跃层的数目越多,学生模型的计算量越大. 式(12)改进之后为

$$\mathcal{L}^2=-\gamma\left[\left(\frac{1}{\sum q_i(x)\lg\left(\frac{q_i(x)}{p_i(x)}\right)}\right)\left(\sum_{i=1}^{l}\left(g_i-t\right)\right)\right] \tag{13}$$

但训练过程中容量的动态伸缩会导致学生模型难以收敛,鉴于此,提出针对t的热启动策略:

$$t=\min\left(\frac{E}{w},m\right) \tag{14}$$

约束学生模型在训练早期阶段约束g_i,缓慢增加容量进行蒸馏学习,促进其收敛. 式(14)中,E为当前轮次,m为系数最大值,而w为热启动参数,控制热启动策略作用的时长.

(2)其次,松弛策略在蒸馏学习的后期移除对学生模型容量的约束. 具体来说,在蒸馏学习后期移除了损失项,以鼓励学生模型充分学习教师知识,同时避免容量约束的限制. 尽管这会导致学生模型容量增加,但与其带来的性能增益相比,容量增加所引起的资源消耗可忽略不计. 另一方面,如果过早移除损失项,学生模型的容量将缺乏约束,难以在蒸馏性能与模型容量之间实现均衡.

3.5 模型蒸馏过程

首先,给定当前批次的样本及其对应的师生模型输出分布,低秩知识变换模块通过高效的低秩变换矩阵,将教师模型的输出分布分别变换到类别标签和学生模型的输出分布. 对应的损失项如式(6)所示. 利用该损失项优化变换矩阵待其收敛之后,获取输出 O^1 和 O^2. 然后,伸缩感知模块根据当前蒸馏动态对学生模型架构进行动态构建,优化决策变量来约束学生模型容量,其对应损失项为

$$L=\mathcal{L}_{\mathrm{CE}}+\mathcal{L}_{\mathrm{KL}}(p,q)+\mathcal{L}_{\mathrm{KL}}(O^1,q)+\mathcal{L}_{\mathrm{KL}}(O^2,q)+\mathcal{L}^2 \tag{15}$$

最后,在松弛阶段本文移除 $\mathcal{L}^2$ 损失.

4 实验评估

4.1 数据集

为了验证本文所提方法的有效性,在2个图像通用数据集(CIFAR-100[15]和ImageNet[16])进行了实验. CIFAR-100为公开经典的图像数据集,它由100个不同类别的RGB图像组成,包含5万训练图像和1万测试图像. ImageNet是一个由1 000个不同类别组成的大规模图像数据集,对于知识蒸馏场景下的学生模型来说更具挑战性. ImageNet训练集包含128万张图像,验证集包含5万张图像.

4.2 实验设置

为了公平比较,本文对实验所使用的蒸馏方法进行相同的训练设置,包括数据预处理、学习率调度器、训练周期数和批处理大小等超参数. 本文使用SGD优化对LAFA方法的参数进行优化. 其他超参数设置如表1所示,其中,b 为最小批次大小,lr为初始学习率,p 为学习率发生衰减的轮次,epoch为模型训练轮次,τ 为蒸馏温度,α 和 β 为低秩知识变换模块的损失函数 $\mathcal{L}^1$ 的权重,γ 为伸缩感知模块的损失函数 $\mathcal{L}^2$ 的权重,t 为比例系数,w 为热启动参数. 在蒸馏过程的每一轮次中,分别针对2个低秩矩阵独立优化20轮,其中低秩矩阵是由多层感知机MLP(MultiLayer Perceptron)实现的. 对于师生模型架构,采用主流的ResNet[17]、VGG[18]、ShuffleNet-V1[19]、WideResNet[20]和DenseNet[21]等架构.

4.3 对比实验

4.3.1 CIFAR-100结果

为了检验LAFA方法对学生模型性能的提升效果,本文将其作为一种即插即用的策略集成在对比方法中(表示为"模型-D"),并在8组不同师生模型架构下进行了比较. 结果如表2和表3所示.

表1　本文LAFA方法的超参数设置

	超参设置	b	lr	p	epoch	τ	α	β	γ	t	w
数据集	CIFAR-100	64	0.05	150、180、210	240	4	0.5	0.5	10	0.95	5
	ImageNet	512	0.10	30、60、90	100	4	0.7	0.7	10	0.95	5

表2　蒸馏方法集成LAFA在CIFAR-100数据集的Top-1准确率对比　　单位:%

教师模型	ResNet56 72.34		ResNet56 72.34		ResNet110 74.31		ResNet110 74.31	
学生模型	ResNet20 69.09	ResNet20-D —	ResNet14 67.32	ResNet14-D —	ResNet32 71.14	ResNet32-D —	ResNet44 72.06	ResNet44-D —
KD[1]	70.66	71.40	68.24	68.40	73.08	73.76	74.43	74.51
WSL[22]	71.33	71.62	68.92	69.36	73.14	73.46	74.40	74.42
AnnealingKD[23]	70.04	70.95	67.27	67.77	73.30	73.69	73.97	74.41
DKD[2]	71.97	72.04	67.79	68.57	74.11	74.52	74.61	74.75
ReKD[24]	71.30	71.64	69.15	69.59	73.43	73.57	74.47	74.55
FITNET[7]	69.21	69.57	67.43	67.34	71.06	70.97	71.57	71.53
RKD[25]	69.61	69.78	66.69	67.17	71.82	71.87	72.58	72.79
OFD[26]	70.98	70.53	66.55	66.87	73.23	73.43	73.57	74.04
PKT[27]	70.34	70.91	68.38	69.03	72.21	72.59	73.41	73.56
VID[8]	69.88	70.33	68.03	68.08	72.12	72.25	73.59	73.49
SP[28]	70.33	70.94	67.80	68.14	72.93	72.82	73.80	73.95
REVIEWKD[3]	71.89	71.91	69.48	69.47	73.89	74.03	74.66	74.88
NRKD[4]	71.50	71.67	68.37	68.55	73.89	74.05	74.75	75.15

表3 蒸馏方法集成LAFA在CIFAR-100数据集的Top-1准确率对比 单位:%

教师模型	WRN-40-2 75.61		WRN-40-2 75.61		WRN-40-2 75.61		VGG13 74.64	
学生模型	WRN-40-1 71.98	WRN-40-1-D —	WRN-16-2 73.26	WRN-16-2-D —	ShuffleNet-V1 70.50	ShuffleNet-V1-D —	DenseNet 64.17	DenseNet-D —
KD[1]	73.54	73.79	74.92	75.14	74.83	75.39	66.61	67.05
WSL[22]	73.76	73.82	75.40	75.55	73.19	75.03	65.65	66.03
AnnealingKD[23]	73.65	74.03	75.12	75.40	75.88	76.03	67.24	67.41
DKD[2]	74.81	75.01	76.24	76.02	76.70	76.76	67.69	68.02
ReKD[24]	74.14	74.41	75.51	75.55	76.09	76.51	65.93	66.26
FITNET[7]	72.24	72.33	73.58	73.75	73.73	73.91	64.04	64.28
RKD[25]	72.22	72.39	73.35	73.90	72.21	72.95	64.40	65.23
OFD[26]	74.33	74.42	75.24	75.83	75.85	76.98	65.37	64.47
PKT[27]	73.13	73.21	73.97	74.90	71.18	74.44	65.77	66.00
VID[8]	71.71	71.76	73.11	73.30	71.16	72.04	63.65	64.19
SP[28]	73.33	73.60	73.93	74.40	75.53	75.68	66.92	67.12
REVIEWKD[3]	75.09	74.56	76.12	76.04	75.90	76.32	65.32	65.45
NRKD[4]	74.22	74.88	75.40	76.12	76.84	77.09	67.56	67.66

(1)LAFA方法集成后,均对学生模型的性能有所提升(在0.5~6.5个百分点之间). 这与知识蒸馏对教师模型知识转移的效果一致. 基于逻辑知识(Decoupled Knowledge Distillation, DKD)和基于特征知识(REVIEWKD)的方法相对于其他对比方法对学生模型性能提升较大,提升在2.2~6.0个百分点. 这是因为它们在蒸馏学习中分解监督知识和分别细化特征来蒸馏,挖掘多个知识之间的相关性. 并且,LAFA方法集成在对比方法上,在大多数情况下超过了现有的对比方法(在0.3~1.0个百分点之间). 这得益于低秩知识变换模块对教师模型分布知识多样性的提高和伸缩感知模块对学生模型容量的动态伸缩.

(2)在师生模型容量差方面,实验结果表明,LAFA方法在师生模型容量差距较小的情况下性能较好. 比如在ResNet56和ResNet20-D(容量差距小)中,性能提升在0.2~0.9个百分点,而在ResNet56和ResNet14-D(容量差距较大)中,性能提升在0.2~0.7个百分点. 原因包括:①容量较小的学生模型在蒸馏早期表现较差,低秩变换模块的优化效果受限;②在引入伸缩感知模块后,其模型层数较少(一般10层左右),此时伸缩感知模块容易跳跃一些对性能提升较为关键的中间层.

4.3.2 ImageNet结果

本文在更具挑战性的ImageNet数据集上验证了LAFA方法的有效性,采用ResNet34作为教师模型,ResNet18和ResNet10作为学生模型. 在对比方法上集成LAFA方法并观察性能对比,使用Top-1和Top-5作为评价指标,结果如表4所示. 在绝大部分性能对比实验中,LAFA方法能有效提升现有蒸馏方法在学生模型上的性能,在Top-1准确率上提升在0.10~0.48个百分点之间.

表4 蒸馏方法集成LAFA在ImageNet数据集上的准确率对比

单位:%

性能指标	学生模型	AT[29]	OFD[26]	REVIEWKD[3]	KD[1]	DKD[2]
Top-1	ResNet18	70.69	70.81	71.61	70.66	71.60
	ResNet18-D	70.73	70.66	71.71	71.14	71.73
Top-5	ResNet18	90.01	89.98	90.51	89.88	90.41
	ResNet18-D	89.93	89.78	90.57	90.03	90.44
Top-1	ResNet10	62.59	62.34	64.41	64.14	65.34
	ResNet10-D	62.67	62.37	64.30	64.21	65.67
Top-5	ResNet10	84.48	84.52	86.09	85.59	86.43
	ResNet10-D	84.49	84.41	85.74	85.54	86.44

4.4 消融实验

本节通过消融实验验证LAFA方法中各模块的有效性. 本文构造了如下4个变体方法:(1)L_1,仅配备低秩知识变换模块;(2)L_2,仅配备伸缩感知模块;(3)L_3,配备伸缩感知模块及热启动策略;(4)L_4,配备伸缩感知模块及松弛策略.

表5展示了这4种变体集成在普通KD方法上的性能提升对比. 由表5可知,所有变种方法在不同程度上提升了蒸馏性能. 特别地,伸缩感知模块中的热启动策略在Top-1指标上为学生模型带来了显著提升,ResNet-20和ResNet-32分别提升了约0.7个百分点和0.6个百分点. 这表明,学生模型通过缓慢增加容量进行蒸馏学习,有助于其稳定收敛,从而降低在容量动态伸缩过程中的训练难度,最终提升蒸馏性能.

表5　LAFA方法及其4个变种方法的准确率对比　　单位:%

学生模型	性能指标	KD[1]	L_1	L_2	L_3	L_4
ResNet20	Top-1	70.66	71.04(+0.38)	71.02(+0.36)	71.36(+0.70)	70.81(+0.15)
	Top-5	92.18	92.23(+0.05)	92.25(+0.07)	92.37(+0.19)	92.47(+0.29)
ResNet32	Top-1	73.08	73.31(+0.23)	72.94(−0.14)	73.76(+0.68)	73.33(+0.25)
	Top-5	93.16	93.27(+0.11)	92.96(−0.20)	93.46(+0.30)	93.51(+0.35)

4.5　分析实验

4.5.1　低秩知识变换模块

为验证低秩变换模块中的变换矩阵所生成监督知识的有效性,本文针对不同监督策略进行了性能对比.表6展示了实验结果,Baseline、$\boldsymbol{S}_1$、$\boldsymbol{S}_2$和$\boldsymbol{S}_1$+$\boldsymbol{S}_2$分别表示学生模型采用不同监督策略.Baseline为仅教师输出蒸馏,$\boldsymbol{S}_1$为教师和$\boldsymbol{S}_1$输出共同蒸馏,$\boldsymbol{S}_2$为教师和$\boldsymbol{S}_2$输出共同蒸馏,$\boldsymbol{S}_1$+$\boldsymbol{S}_2$为三者联合蒸馏.结果表明,2个变换矩阵显著提升了蒸馏性能.这表明:(1)即使是过参数的教师模型,蒸馏性能仍有提升空间,且通过变换矩阵能够实现这一提升;(2)引导教师模型逆向学习学生模型能够生成更适合学生模型的知识.传统知识蒸馏方法忽视了师生模型之间的差异,弥补这些差异有助于提升蒸馏学习效果.此外,2个变换矩阵的提升效果不同,$\boldsymbol{S}_1$矩阵对学生模型的准确率提升了0.7个百分点,而$\boldsymbol{S}_2$矩阵使准确率提升了1.02个百分点.这一差异表明,在训练初期,学生模型性能较差,$\boldsymbol{S}_1$矩阵生成的知识不够准确,导致性能提升受限.因此,在蒸馏前期应避免过度依赖学生模型输出构建的知识,而应在学生模型性能提升后再进行构建.

表6　不同变换矩阵监督下的学生模型准确率　　单位:%

监督类型	Baseline	S_1	S_2	S_1+S_2
Top-1	72.50	73.20	73.52	73.21
Top-5	93.10	93.24	93.21	93.08

4.5.2　伸缩感知模块

本文设计了3个实验分别分析决策变量、热启动和松弛策略的有效性.

(1)通过优化比例系数t约束决策变量.该系数决定了师生模型之间的容量差距.为分析该系数对蒸馏学习的影响,设置了不同的t值并记录其对应蒸馏性能.结果如图2所示,一方面,较小的学生模型容量导致师生容量差距过大,从而使学生难以有效蒸馏知识(如ResNet20在t=0.1时,准确率仅为56.97%).另一方面,并非学生模型容量越大性能越好,3个学生模型在t趋向于1时,性能先达到峰值然后有所降低.这是因为,学生模型在面对较简单样本来说,过度利用其全部容量进行学习容易导致过拟合,并失去了感知蒸馏知识难度的能力,无法稳定提升性能.最后,t接近0.85时,学生模型的性能达到峰值,其模型容量适配教师模型,表明决策变量促进了学生模型容量和蒸馏性能达到平衡.

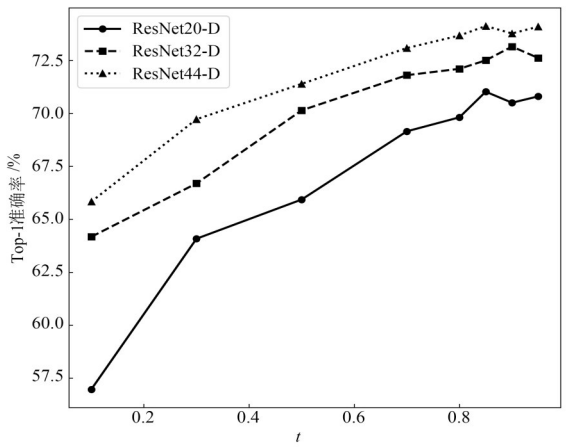
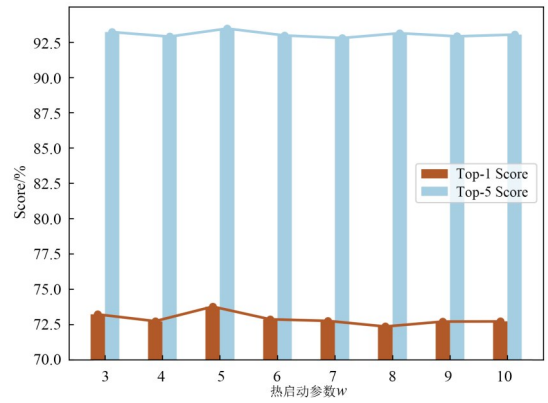


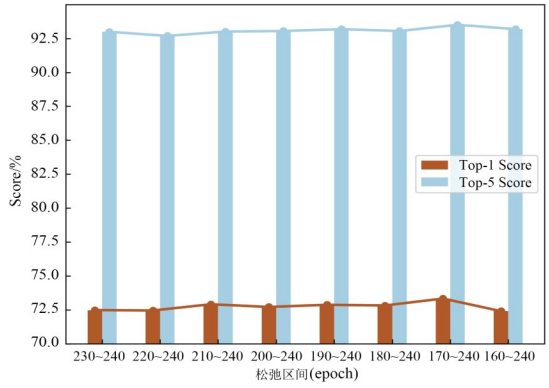
图2　3种学生模型在不同比例下的准确率趋势图

(2)为分析热启动策略的有效性,设置了不同的参数w并观察对应的学生模型性能.实验结果如图3(a)所示,随着w值增大,学生模型的性能首先达到峰值然后下降.通过在训练早期跳过某些层,减少模型对训练数据的过度拟合,促使网络学习到更具泛化性的特征表示.随着训练的进行,跳跃频率逐渐降低,从而在优化模型容量的同时保持性能.然而,w值过大,可能会导致学生模型容量在训练早期不足且增长过慢,难以学习蒸馏教师模型的复杂知识,导致学生模型性能降低;相反,w值过小则可能导致学生模型过早依赖自身容量进行学习,从而导致过拟合;当w=5时,学生模型能够在蒸馏性能和容量之间达到平衡,平稳地达到最优性能.

(3)为验证松弛策略的有效性,设置了不同松弛区间并观察对应的学生模型性能.如图3(b)所示,在170~240轮次实施松弛策略,学生模型性能达到最优,在Top-1上获得了0.8个百分点的性能提升.如果过早实施松弛策略,学生模型容易趋向于使用全部的容量来蒸馏学习,造成模型过拟合的现象.另一方面,过晚实施松弛策略,学生模型缺乏足够的时间增加模型容量,导致后期松弛区间的性能提升不足,难以达到性能和容量的均衡.综上所述,合理选择松弛策略的实施时



(*a*) 热启动参数 w



(*b*) 松弛区间

图3 LAFA方法参数敏感度分析

机对于优化模型性能至关重要.

5 结论与展望

本文提出了一种低秩自适应伸缩感知蒸馏方法LAFA. 其通过构建高效的低秩变换矩阵,将教师模型的输出分布分别变换到学生模型输出分布和类别标签,以提高分布知识的多样性. 同时,基于决策辅助器对学生模型架构进行动态构建,并根据当前师生模型之间的输出分布差距伸缩模型容量. 特别地,本文提出热启动和松弛策略来优化决策变量,缓解了学生模型难以收敛的问题,并通过少量资源消耗实现了显著性能提升. 在2个图像分类数据集上进行了对比实验、消融实验和分析实验,实验结果表明,LAFA方法具有有效性.

本文将来的工作主要包括:(1)针对LAFA方法中的低秩变换矩阵,将考虑设计更为高效的变换工具(如图神经网络和注意力模型等)对特征知识等进行变换;(2)探索LAFA方法在其他领域的应用潜力. 例如,在多任务学习场景下,为每个任务构建低秩变换矩阵,并根据任务相关性动态分配权重,以优化模型在多任务学习中的表现.

参考文献

[1] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2024-10-07]. https://arxiv.org/abs/1503.02531v1.

[2] ZHAO B R, CUI Q, SONG R J, et al. Decoupled knowledge distillation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 11943-11952.

[3] CHEN P G, LIU S, ZHAO H S, et al. Distilling knowledge via knowledge review[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 5006-5015.

[4] XIN X M, SONG H P, GOU J P. A new similarity-based relational knowledge distillation method[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2024: 3535-3539.

[5] MIRZADEH S I, FARAJTABAR M, LI A, et al. Improved knowledge distillation *via* teacher assistant[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 5191-5198.

[6] LIU Y, JIA X H, TAN M X, et al. Search to distill: Pearls are everywhere but not the eyes[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 7536-7545.

[7] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: Hints for thin deep nets[EB/OL]. (2015-03-27)[2024-10-07]. https://arxiv.org/abs/1412.6550v4.

[8] AHN S, HU S X, DAMIANOU A, et al. Variational information distillation for knowledge transfer[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 9163-9171.

[9] YOU S, XU C, XU C, et al. Learning from multiple teacher networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 1285-1294.

[10] 欧阳毅, 汤文燕, 黎晏伶. 基于特征蒸馏的变分编码器交通流预测模型[J]. 电子学报, 2024, 52(6): 1938-1944.
OUYANG Y, TANG W Y, LI Y L. Traffic flow prediction model based on spatio-temporal feature distillation

variational autoencoder[J]. Acta Electronica Sinica, 2024, 52(6): 1938-1944. (in Chinese)

[11] 郑云飞, 王晓兵, 张雄伟, 等. 基于金字塔知识的自蒸馏HRNet目标分割方法[J]. 电子学报, 2023, 51(3): 746-756.
ZHENG Y F, WANG X B, ZHANG X W, et al. The self-distillation HRNet object segmentation based on the pyramid knowledge[J]. Acta Electronica Sinica, 2023, 51(3): 746-756. (in Chinese)

[12] WANG X, YU F, DOU Z Y, et al. SkipNet: Learning dynamic routing in convolutional networks[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 420-436.

[13] SAINATH T N, KINGSBURY B, SINDHWANI V, et al. Low-rank matrix factorization for Deep Neural Network training with high-dimensional output targets[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013: 6655-6659.

[14] HUIJBEN I A M, KOOL W, PAULUS M B, et al. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1353-1371.

[15] KRIZHEVSKY A. Learning multiple layers of features from tiny images[EB/OL]. (2009-04-08) [2024-10-07]. https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.

[16] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2009: 248-255.

[17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[18] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2024-10-07]. https://export.arxiv.org/abs/1409.1556v6.

[19] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6848-6856.

[20] ZAGORUYKO S, KOMODAKIS N. Wide residual networks[C]//Proceedings ofthe British Machine Vision Conference 2016. Paris: British Machine Vision Association, 2016.

[21] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2261-2269.

[22] ZHOU H L, SONG L C, CHEN J J, et al. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective[EB/OL]. (2021-02-01) [2024-10-07]. https://arxiv.org/abs/2102.00650.

[23] JAFARI A, REZAGHOLIZADEH M, SHARMA P, et al. Annealing knowledge distillation[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. EACL: Association for Computational Linguistics, 2021: 2493-2504.

[24] XU L C, REN J, HUANG Z H, et al. Improving knowledge distillation via head and tail categories[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(5): 3465-3480.

[25] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 3962-3971.

[26] HEO B, KIM J, YUN S, et al. A comprehensive overhaul of feature distillation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1921-1930.

[27] PASSALIS N, TEFAS A. Learning deep representations with probabilistic knowledge transfer[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 283-299.

[28] TUNG F, MORI G. Similarity-preserving knowledge distillation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1365-1374.

[29] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[EB/OL]. (2017-02-12)[2024-10-07]. https://arxiv.org/abs/1612.03928v3.

作者简介

李佳明 男,2002 年 10 月生,广东梅州人. 硕士研究生. 主要研究方向为知识蒸馏和计算机视觉.
E-mail: lijm071@163.com

鲍志强 男,1995 年 1 月生,江西九江人. 博士. 主要研究方向为知识蒸馏和模型压缩.
E-mail: zhiqiangbao1995@163.com

黄震华 男,1980 年 9 月生,福建莆田人. 教授,博士生导师. 主要研究方向为机器学习、数据挖掘、推荐系统.
E-mail: jukiehuang@163.com

孙圣力 男,1978 年 12 月生,湖南常德人. 博士,教授. 主要研究方向为机器学习、数据挖掘、数据库.
E-mail: slsun@ss.pku.edu.cn

陈运文 男,1981 年 7 月生,江苏南京人. 博士,高级工程师. 主要研究方向为机器学习、数据挖掘、自然语言处理.
E-mail: chenyunwen@datagrand.com