软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2414−2429 [doi: 10.13328/j.cnki.jos.006958]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向小样本学习的轻量化知识蒸馏*

陈嘉言, 任东东, 李文斌, 霍 静, 高 阳

(南京大学 计算机科学与技术系, 江苏 南京 210023)

通信作者: 李文斌, E-mail: liwenbin@nju.edu.cn

摘 要: 小样本学习旨在模拟人类基于少数样例快速学习新事物的能力, 对解决样本匮乏情境下的深度学习任务具有重要意义. 但是, 在诸多计算资源有限的现实任务中, 模型规模仍可能限制小样本学习的广泛应用. 这对面向小样本学习的轻量化任务提出了现实的需求. 知识蒸馏作为深度学习领域广泛使用的辅助策略, 通过额外的监督信息实现模型间知识迁移, 在提升模型精度和压缩模型规模方面都有实际应用. 首先验证知识蒸馏策略在小样本学习模型轻量化中的有效性. 并结合小样本学习任务的特点, 针对性地设计两种新的小样本蒸馏方法: (1) 基于图像局部特征的蒸馏方法; (2) 基于辅助分类器的蒸馏方法. 在 miniImageNet 和 TieredImageNet 数据集上的相关实验证明所设计的新的蒸馏方法相较于传统知识蒸馏在小样本学习任务上具有显著优越性.

关键词: 深度学习; 小样本学习; 图像识别; 知识蒸馏; 模型轻量化

中图法分类号: TP18

中文引用格式: 陈嘉言, 任东东, 李文斌, 霍静, 高阳. 面向小样本学习的轻量化知识蒸馏. 软件学报, 2024, 35(5): 2414–2429. http://www.jos.org.cn/1000-9825/6958.htm

英文引用格式: Chen JY, Ren DD, Li WB, Huo J, Gao Y. Lightweight Knowledge Distillation for Few-shot Learning. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2414–2429 (in Chinese). http://www.jos.org.cn/1000-9825/6958.htm

Lightweight Knowledge Distillation for Few-shot Learning

CHEN Jia-Yan, REN Dong-Dong, LI Wen-Bin, HUO Jing, GAO Yang

(Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China)

Abstract: Few-shot learning aims at simulating the ability of human beings to quickly learn new things with only few samples, which is of great significance for deep learning tasks when samples are limited. However, in many practical tasks with limited computing resources, the model scale may still limit a wider application of few-shot learning. This study presents a realistic requirement for lightweight tasks for few-shot learning. As a widely used auxiliary strategy in deep learning, knowledge distillation transfers knowledge between models by using additional supervised information, which has practical application in both improving model accuracy and reducing model scale. This study first verifies the effectiveness of the knowledge distillation strategy in model lightweight for few-shot learning. Then according to the characteristics of few-shot learning, two new distillation methods for few-shot learning are designed: (1) distillation based on image local features; (2) distillation based on auxiliary classifiers. Experiments on miniImageNet and TieredImageNet datasets demonstrate that the new distillation methods are significantly superior to traditional knowledge distillation in few-shot learning tasks.

Key words: deep learning; few-shot learning; image recognition; knowledge distillation (KD); model lightweight

知识蒸馏 (knowledge distillation, KD) 的基本思想是采用一个训练好的深度网络 (称为“教师模型”) 去辅助训练一个新的深度网络 (称为“学生模型”). 学生模型在训练过程中, 会同时接收来自真实样本标签和教师模型输出两方面的监督, 综合两种监督信息对自身的输出进行修正. 知识蒸馏策略被广泛应用于模型轻量化任务, 使用较大

* 基金项目: 国家自然科学基金 (62106100, 62192783, 62276128); 江苏省自然科学基金 (BK20221441); 江苏省双创博士项目 (JSSCBS20210021)
收稿时间: 2022-11-08; 修改时间: 2022-12-15, 2023-03-23; 采用时间: 2023-04-13; jos 在线出版时间: 2023-09-27
CNKI 网络首发时间: 2023-10-07

参数规模的教师模型辅助训练较小参数规模的学生模型. 相较于缩减层数, 通道剪枝等直接减少网络参数的轻量化方法, 知识蒸馏通过重训练实现师生间的知识传递, 更有利于保持原始模型的性能.

小样本学习 (few-shot learning) 旨在模拟人类基于已有知识快速学习新事物的能力, 使模型能在具有一定预知识的基础上, 只通过单个或少数几个样本快速掌握新的未见类别. 在小样本学习任务中, 用作学习预知识的训练集与用作检验新类别学习能力的测试集类别彼此互斥, 且每个新类别只包含极少量可见数据. 这样的任务设置大幅提升了学习任务本身的难度.

样本量的匮乏导致小样本学习模型难以准确掌握目标类别的特征分布, 对模型的训练方法提出了更高的要求. 知识蒸馏策略凭借其高效的知识传递能力, 成为小样本学习领域一种重要的辅助技巧. 围绕知识蒸馏策略与小样本学习方法的结合, 此前已有研究者进行过一定程度的探索, 但仍存在明显的局限性: 首先, 此前的相关工作大多以进一步发掘小样本学习模型本身的性能为目的, 并未深入探索知识蒸馏对于小样本模型轻量化的效果; 其次, 此前的工作中知识蒸馏只作用于模型的预训练阶段, 与核心的小样本任务关联性较弱, 学生模型的训练流程仍然需要经过“预训练-小样本微调”两个阶段, 实际应用中欠缺灵活性.

针对上述问题, 本文重点探究了知识蒸馏策略在小样本学习模型轻量化中的应用. 轻量化的学生模型直接基于小样本任务从头开始训练, 在教师模型输出和真实样本标签的联合指导下进行模型优化, 证明了知识蒸馏策略的有效性. 此外, 针对小样本任务中单个子任务所含类别数较少且随机组合的特点, 本文设计了两类小样本蒸馏方法.

(1) 基于图像局部特征的蒸馏方法 (local feature distillation, LFD): 利用特征提取网络产生的图像中间层特征提供辅助监督信息.

(2) 基于辅助分类器的蒸馏方法 (auxiliary classifier distillation, ACD): 利用附加线性分类器产生的固定分类概率提供辅助监督信息.

在 miniImageNet 和 TieredImageNet 数据集上的相关实验证明, 本文设计的新蒸馏方法相较于传统知识蒸馏方法, 能给小样本学习任务中的轻量化学生模型带来更显著的性能提升. 源代码请见 https://github.com/cjy97/FSLKD

本文基于计算机视觉领域经典的图像分类识别任务展开说明. 第 1 节具体介绍知识蒸馏和小样本学习的基本概念和传统实现方法. 第 2 节围绕“面向小样本学习的轻量化知识蒸馏”这一主题, 详细说明本文的任务设置, 包括选用的特征提取网络结构, 小样本任务的分类器, 学生模型的产生方式等. 第 3 节重点介绍本文设计的新蒸馏方法, 并具体分析其在小样本学习任务中相较于传统蒸馏方法的优势. 第 4 节为实验结果, 展示知识蒸馏策略在小样本学习模型轻量化任务中的有效性, 结合对比实验证明新蒸馏方法的优势. 第 5 节为总结与展望.

1 背景介绍

1.1 知识蒸馏

知识蒸馏采用“师-生模型”[1]训练模式, 通过一个预训练好的教师模型的输出, 辅助新的学生模型训练. 教师模型本身在学生模型的训练过程中通常保持不变. 根据所用学生模型产生方式及任务目的的不同, 知识蒸馏策略可以大致分为“自蒸馏”[2,3]和“轻量化蒸馏”[4,5]两类. 自蒸馏任务中, 学生模型采用与教师模型相同的结构, 并要求其训练后的性能可以超过教师模型, 否则蒸馏无意义. 轻量化蒸馏任务中, 学生模型的参数规模要大幅小于教师模型. 此时不强制要求学生模型训练后的性能超过教师, 只要优于其在无蒸馏训练时的对照性能, 就可以证明蒸馏策略的有效性.

传统基于深度学习的图像分类识别任务[6]中, 对于每张输入的样本图像, 深度网络模型会产生一组分类概率. 分类概率的长度等于训练数据集所含的类别总数. 所得分类概率会与样本图像的真实类别标签 (label, 通常先转化为独热编码形式) 进行对比, 经损失函数计算得到训练损失, 通过梯度回传更新深度网络的权重. 分类识别任务中最常用的损失函数是交叉熵 (cross entropy, CE), 用于度量分类概率与真实标签之间的差异程度.

在引入了知识蒸馏策略的分类识别任务中, 每张输入样本图像会分别通过学生和教师模型, 得到两组同维的

分类概率, 记作 s 和 t. 如果将独热编码形式的真实标签称为“硬标签” (只有类别位是 1, 其余都是 0), 那么教师模型的输出概率可以视作是一种“软标签” (每一位概率值都介于 0–1 之间, 类别越接近概率值越大), 为学生模型的输出提供另一种额外约束信息. 相较于独热编码硬标签在结果约束中“非对即错”的绝对性, 这种软标签对各个类别都给予一定的置信度, 更有助于从多角度辅助学生模型的训练.

在图像分类识别任务中, 最常用的蒸馏损失定义是 KL 散度 (Kullback-Leibler divergence)[1], 该损失基于师生模型分别所得的分类概率 t 和 s 进行计算, 可以量化反映两组同维概率分布之间的差异程度, 其公式化表述如公式 (1). 公式中 x 表示独立事件, 在分类识别任务中就是当前样本属于每一个类别的情况. $P_S(x)$ 和 $P_T(x)$ 分别表示学生模型和教师模型在 x 事件上的概率, 即模型预测的分类概率在对应位的激活值. log() 为自然对数函数.

$$\begin{aligned} D_{KL}(s_i|t_i) &= \sum_x P_T(x)\log\left(\frac{P_T(x)}{P_S(x)}\right) \\ &= \sum_x P_T(x)\log(P_T(x)) - \sum_x P_T(x)\log(P_S(x)) \end{aligned} \tag{1}$$

除了传统的基于分类概率的蒸馏方法, 还有一类蒸馏方法利用特征提取网络所得的图像中间层特征或特征间关系进行约束. 此类工作的基本思想是: 对于同一输入样本, 理想的学生模型与教师模型所产生的特征分布也应该是相似的. 因此图像特征本身也可以指导学生模型的训练, 使其产生的特征分布去拟合教师模型的输出. Romero 等人[7]最早尝试基于网络中间层特征进行知识蒸馏, 并初步设计了多层特征间约束的具体方案; Zagoruyko 等人[8]进一步提出了对师生模型各自所得的注意力图进行约束, 让学生模型学习的特征更加灵活多样; Tung 等人[9]和 Park 等人[10]的工作则摆脱对特征值本身的依赖, 转而靠特征间的相似性关系作为蒸馏凭据, 促使学生模型学习教师模型提炼数据内部关系的能力.

知识蒸馏策略还被广泛应用于模型轻量化任务. 蒸馏方法的应用只要求学生和教师模型能产生相同形式的输出, 对二者的具体内部结构并没有强制性的限制. 因此轻量化蒸馏任务中会采用参数规模相对小的网络作为学生模型, 以降低后续使用时的模型推理成本. Liu 等人[11]的工作规范了轻量化知识蒸馏的问题定义, 并从损失入手改进了传统的蒸馏方式; Hou 等人[12]的工作借鉴自蒸馏方法, 直接训练轻量级模型完成专一化视觉任务. 理想情况下, 轻量化蒸馏能够让学生模型以远少于教师模型的参数量, 取得与完整教师相差无几的性能.

轻量化的学生模型有多种产生方式 (见图 1). 一类方法是直接选择参数量少于教师模型的其他结构网络作为学生. 另一类方法则是在完整教师模型的基础上, 通过缩减层数, 通道剪枝[13]等方法, 间接得到轻量化的模型. 相比之下, 后者所得的学生模型在基本网络结构上与原教师模型更为相似, 在相同条件下通常更容易取得良好的蒸馏训练效果.

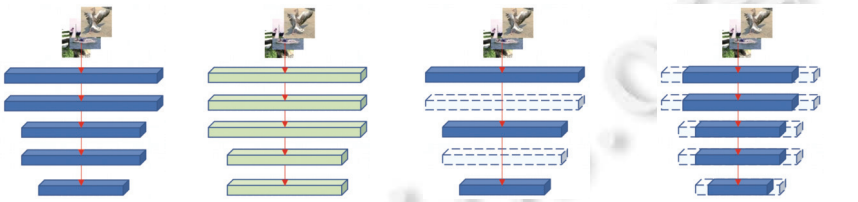


图 1 轻量化模型产生方式

1.2 小样本学习

小样本学习[14]旨在仅利用少量样本使模型快速学习新的类别. 样本量的匮乏导致可供学习的特征分布过分单一而缺乏普适性, 给小样本学习任务带来了更高的难度. 相较于传统深度学习, 小样本学习在任务设置上存在一个显著区别: 训练集与测试集的类别空间彼此互斥. 此时, 用于小样本学习任务的深度模型的学习目标不再是训练集类别本身的特征, 而是根据少量数据快速学习新类别的能力. 这种高层次的学习能力与任一特定的图像类别无关, 因此能被迁移[15]到新的未见类别上, 使得小样本模型能够通过少量的新样本, 快速学习此前没有见过的全新

的样本类.

为了构造小样本的任务情境, 常遵循“episodic training”模式[16,17]对小样本模型进行训练. 整个训练过程被拆分成若干子任务 (episode). 每个子任务包含 N 个训练集类别, 每个类别又分别包含 K 个支持集样本 (support samples) 和若干查询集样本 (query samples), 称作一个“N-way K-shot”的子任务. 其中, 支持集样本供模型快速学习当前子任务所含 N 个类别的特征. 模型据此对查询集样本进行分类预测, 所得结果用于验证精度和计算训练损失. 测试阶段也采用类似的方式, 只是子任务抽样的对象变为测试集所含类别. 如此, 在每一个子任务中, 小样本分类器都只进行 N 个类别上的分类预测. 每个子任务所含的类别组合是完全随机的.

在上述任务设置下, 小样本学习模型在每个子任务中的处理流程, 大致可以划分为特征提取和度量分类两个阶段. 特征提取阶段和传统深度学习基本相同, 图像特征提取网络通常为卷积神经网络 (convolutional neural network, CNN) 结构. 样本图像输入特征提取网络, 经过若干卷积, 池化和激活函数的处理, 得到最终的特征表示. 根据特征提取网络末尾是否包含全局均值池化层 (global average pooling, *GAP*), 所得的图像特征又可分为全局特征 (使用 *GAP* 层, 每张图像产生单个特征向量) 和局部特征 (不使用 *GAP* 层, 每张图像产生一组 n>1 个特征向量). 两类特征如图 2 所示, 其公式化表述如公式 (2) 和公式 (3), 其中 f 表示特征提取网络.

$$F_{\text{local}}=f(x)=[t_1,t_2,\ldots,t_n]\in\mathbb{R}^{d\times n} \tag{2}$$

$$F_{\text{global}}=GAP(f(x))\in\mathbb{R}^{d} \tag{3}$$

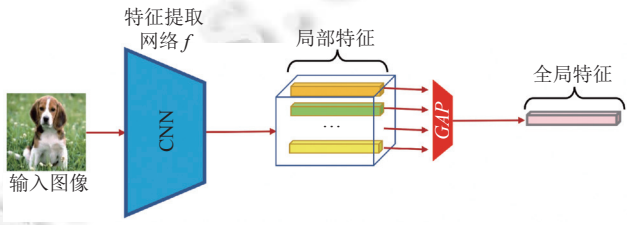


图 2 全局特征对比局部特征

度量分类阶段, 小样本分类器会先对支持集样本所得特征进行处理, 得到当前子任务 N 个类别各自的特征中心表示. 然后, 对于每一个待分类的查询集样本, 分别度量其特征与 N 个类别中心之间的距离或相似度, 从而预测其最有可能属于哪一个类别. 例如, 经典的基于全局特征的原型网络 (prototypical network) 度量分类器[16]中, 直接将每个类别所含支持集样本的全局特征的均值作为类别中心, 然后计算每个查询集样本特征与各个类别中心的距离, 把相距最近的类别作为查询集样本的分类预测结果.

除了直接按照上述“episodic training”方式训练模型的方法外, 还有部分小样本学习方法[18]会在小样本任务训练之前, 先按照传统深度学习分类方法对模型进行预训练 , 以获取一个性能相对理想的特征提取网络. 预训练结束后, 模型移除传统线性分类器, 换用小样本度量分类器, 再按照小样本“episodic training”方式进行微调. 模型经预训练所得的特征提取能力同样能够被不同的样本类别 (包含之后可能遇到的新类别) 所通用, 从而提升后续小样本任务微调的效果. 但这种“预训练-小样本微调”的两阶段训练模式, 也相应增大了小样本学习任务训练的成本和难度.

1.3 知识蒸馏与小样本学习的结合

与传统深度学习任务一样, 小样本学习也可能面临现实应用中计算资源受限的困境. 知识蒸馏策略作为深度学习领域提升模型精度和实现模型轻量化的常用辅助技巧, 在小样本学习领域的有着巨大的应用潜力. 此前已有部分工作尝试将知识蒸馏策略引入到小样本学习任务中. Tian 等人[18]将传统的知识蒸馏方法作用于小样本学习任务的预训练阶段, 使用常规分类交叉熵损失和 KL 散度损失共同训练特征提取网络, 提升模型基础性能; Rajasegaran 等人[19]将预训练进一步划分为两个阶段, 在获取初代模型后, 再应用知识蒸馏策略获取性能更优的二代模型. 上述方法都在一定程度上提升了小样本学习模型的精度.

但是, 现有的知识蒸馏与小样本学习结合的研究仍存在明显局限性. 一方面此前围绕小样本学习知识蒸馏的

相关工作中, 学生模型大多采用了“预训练-小样本微调”的两阶段训练模式. 蒸馏损失只作用于基于常规分类任务进行的模型预训练阶段, 与“*N*-way *K*-shot”设置下的小样本子任务缺乏直接联系. 使用的蒸馏方法仍然是传统的基于常规分类概率计算 KL 散度, 并没有结合小样本学习任务的特点在蒸馏方法上进行创新. 另一方面, 此前的工作大多采用模型自蒸馏设置, 并未探索过知识蒸馏在小样本模型轻量化中的作用. 这在一定程度上限制了知识蒸馏方法在小样本学习领域的进一步应用.

针对上述问题, 本文将从传统蒸馏方法在小样本学习任务中的局限性入手, 从方法层面对小样本知识蒸馏进行改进, 并直接作用于小样本子任务本身. 同时为了探究知识蒸馏对于资源受限条件下小样本学习任务的效果, 将重点围绕轻量化设置开展实验.

2 任务设置

本文面向小样本学习的轻量化知识蒸馏, 进行具体任务设置. 本节将从选用的特征提取网络, 选用的小样本度量分类器, 教师-学生模型的具体产生方式 3 个方面进行详细说明.

2.1 特征提取网络

统一使用小样本学习任务中常用的 12 层残差网络 (12-layer residual network, ResNet12)[20]用于图像特征提取. 在 ResNet12 的原始设计中, 网络末尾包含全局均值池化层, 对每张输入图像得到单个全局特征向量. 本文的实验中, 为了对接使用图像局部特征的小样本度量分类器, 网络末尾的全局池化层被统一移除. ResNet12 特征提取网络会对每张输入图像生成固定数量的一组局部特征集合.

2.2 小样本度量分类器

特征提取网络之后, 使用深度最近邻神经网络 (deep nearest neighbor neural network, DN4) [21]作为小样本度量分类器, 完成 *N*-way *K*-shot 的小样本分类任务. DN4 是一种典型的基于图像局部特征的度量分类器, 利用每个待分类的查询集样本所含特征与支持集特征之间的 *k* 近邻相似度加和作为度量分类的评估指标. 相较于使用全局特征的度量分类方法, DN4 分类器更能充分发掘图像内部的特征信息, 在极其有限的支持集样本上更准确地表征类别中心.

假设特征提取网络 f 对每张输入图像产生 n 个局部特征向量. 每个“*N*-way *K*-shot”小样本任务中抽取的 $N\times K$ 张支持集样本通过特征提取网络 f, 得到每个类别的支持集特征 c; 每张待分类的查询集图像样本 q 也通过 f, 得到包含 n 个局部特征的集合 $f(q)$. 将 $f(q)$ 中的 n 个局部特征与每个类别特征集 c 中所含的 $n\times K$ 个支持集特征两两计算余弦相似度 (cosine similarity, cos). 如公式 (4) 所示, 根据所得余弦相似度的大小对 c 中所含特征重新排序, 使得对任意 $1\leqslant i\leqslant n$ 及 $1\leqslant j_1$, $j_2\leqslant n\times K$, 如果 $j_1<j_2$, 则必有 $\cos(x_i,x_i^{j_1})\geqslant\cos(x_i,x_i^{j_2})$. 在此基础上, 度量分类函数 Φ 的定义如公式 (5): 将相似度数值最高的前 k 项加和, 作为当前局部特征 x_i 关于类别 c 的相似度; 再把所有特征的相似度值加和, 作为当前样本图像 q 与类别 c 的度量评分. 所有类别的评分经过 Softmax 激活函数, 得到最终的小样本度量分类概率.

$$\begin{cases} f(q)=[x_1,x_2,\ldots,x_n],c=\left[x_i^1,x_i^2,\ldots,x_i^{n\times K}\right] \\ j_1<j_2\rightarrow\cos\left(x_i,x_i^{j_1}\right)>\cos\left(x_i,x_i^{j_2}\right) \end{cases} \tag{4}$$

$$\begin{cases} \Phi(f(q),c)=\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{k}\cos\left(x_i,x_i^j\right) \\ \cos(x,y)=\dfrac{x^{\mathrm{T}}y}{\|x\|\cdot\|y\|} \end{cases} \tag{5}$$

此外, DN4 作为一种无参的度量分类器, 其本身并不包含任何可训练参数, 因此可以直接套接在不含全局池化层的特征提取网络之后完成小样本分类任务. 这意味着, 使用 DN4 的小样本分类模型的性能, 只取决于图像特征提取网络本身的特征表达能力.

2.3 教师-学生模型产生方式

本文的实验中, 完整的残差网络 ResNet12 被用作教师模型. 在此基础上, 对教师模型逐层执行通道剪枝, 将所得的轻量化模型 (记作 Slim-ResNet12) 作为学生模型. 剪枝后学生模型的参数会重新随机初始化, 以保证其是从头开始训练的.

教师模型事先已经在数据集的训练子集上, 按照传统的全局特征提取+线性分类器的形式进行预训练. 常规小样本学习训练流程中, 会在预训练所得特征提取网络的基础上, 套接小样本度量分类器进行小样本微调训练, 以进一步提升模型在小样本任务上的精度. 而在本文中, 蒸馏方法只需要教师模型提供图像的中间层特征, 因此直接选取预训练后 (未经小样本微调) 的权重用于初始化教师模型. 作为对比, 教师模型"预训练后直接进行小样本测试"所得精度及"预训练后进一步小样本微调"所得精度, 会分别在后文实验部分展示.

知识蒸馏实验的训练过程中, 学生模型从头开始按照"episodic training"模式进行小样本任务训练, 同时利用自身小样本分类所得交叉熵损失及与教师模型联合计算的蒸馏损失进行优化. 这一过程中教师模型参数固定不变. 关于小样本任务中蒸馏损失的具体计算方法, 将在第 3 节详细分析.

3 小样本蒸馏方法

3.1 传统蒸馏方法在小样本学习中的局限性

传统蒸馏方法直接将基于分类概率的 KL 散度损失应用于小样本学习任务, 流程如图 3 所示. 教师模型在预训练阶段使用的线性分类器 (fully connected layer, 即全连接层, 图 3 中用 fc 标识) 被废弃, 用于提取图像全局特征的 *GAP* 层也被移除, 转而像学生模型一样在末尾套接一个 DN4 度量分类器, 用于完成小样本分类任务. 此时, 输入样本分别经过教师和学生模型的特征提取网络, 得到两组局部特征, 再分别输入给 DN4 度量分类器, 得到子任务 N 个类别上的小样本分类概率. 学生模型所得的小样本分类概率与真实样本标签之间正常计算分类交叉熵损失, 师生模型所得的两组同维分类概率之间按照公式 (1) 基于 KL 散度计算蒸馏损失.

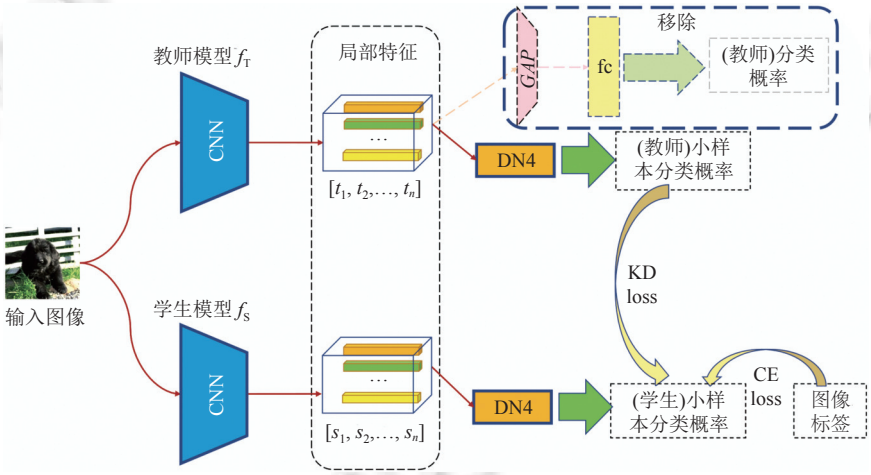


图 3 传统蒸馏方法 (KD) 应用小样本学习任务图示

传统蒸馏方法的应用简单直观, 但由于小样本学习自身在任务设定上的特殊性, 直接将二者结合难以发挥最理想的效果. 如图 4 所示, 左侧显示任务的类别组合, 右侧对应展示蒸馏训练中每个任务在学生模型上产生的梯度更新效果. "参数空间"区域内, t 表示教师模型的参数位置, s_0 表示学生模型的初始参数位置, s_i $(i>0)$ 表示任务 i 对学生模型更新后的参数位置. 箭头指向即梯度更新的方向, 根据先后顺序与每个任务一一对应. 显然, 理想的知识蒸馏应当用尽可能少的任务数, 使得学生模型趋近教师模型的参数分布. 那么梯度箭头就应该尽量指向教师模

型所在的参数位置. 在常规图像分类任务中, 每轮抽样所含样本的类别空间都是固定的, 类别的顺序也恒定不变. 这种性质有利于为学生模型产生稳定的更新方向, 使之尽快拟合教师模型的分布.

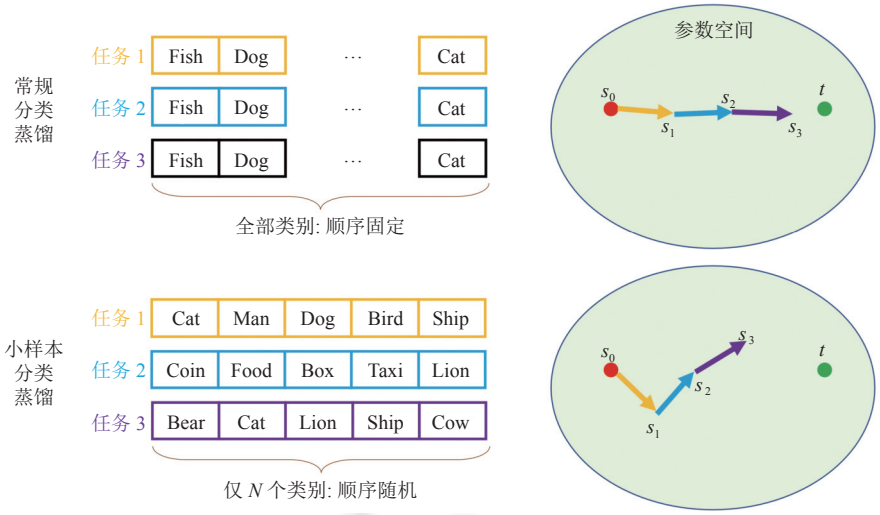


图 4 常规分类蒸馏对比小样本分类蒸馏

相较于传统深度学习分类识别任务, 小样本任务具有两点显著的区别: 一是每个小样本子任务所含的类别数相对很少 (常用设置 $N=5$), 二是每个子任务的类别均为随机抽样产生, 且顺序具有不确定性. 图 4 下半部分, 直观展示了小样本任务中子任务所含类别数较少且随机组合的特点. 这导致传统的基于分类概率的蒸馏方法在小样本任务中存在明显局限性: 每次计算产生的 KL 散度蒸馏损失, 都只适应于当前子任务所产生的类别组合, 其通过梯度回传对模型参数产生的更新, 可能未必有益于后续子任务中随机生成的其他类别组合, 也难以准确指向教师模型对应的梯度方向. 如此, 蒸馏损失对小样本模型的更新变成了一种面向随机子任务亦步亦趋的调节模式, 导致学生模型缺乏一种长期恒定的更新目标, 也因此更难趋近教师模型的性能.

针对上述问题, 本文尝试从其他角度切入, 设计了两类新的专门面向小样本学习任务的蒸馏方法, 能够更充分地利用教师模型已有知识, 深入发掘学生模型的潜在性能.

3.2 基于图像局部特征的蒸馏方法

此前关于知识蒸馏的部分研究[7,8], 已经证明特征提取网络所得的图像中间层特征也可以作为师生模型间蒸馏约束的凭据. 图像中间层特征, 尤其是未经全局池化的局部特征的分布不再与图像的类别强相关, 而是所有类别间通用的性质. 由此, 本文将特征间约束的蒸馏损失应用到图像局部特征上, 提出了针对小样本学习任务的局部特征蒸馏 (local feature distillation, LFD).

局部特征蒸馏方法的基本流程如图 5 所示. 每张样本图像分别通过师生模型的特征提取网络, 得到一组 n 个特征向量的序列, 其中每个特征向量都是对应着原图一块区域的局部特征. 将教师模型的输出特征记作 $[t_1,t_2,\ldots,t_n]$, 学生模型的输出特征记作 $[s_1,s_2,\ldots,s_n]$. 对于任意 $1\leqslant i\leqslant n$, t_i 与 s_i 都对应着同一样本图像的同一区域, 其理论分布应该彼此接近. 局部特征蒸馏中, 让每个局部特征向量分别通过 Softmax 激活函数后, 逐个位置一一对应计算差异程度. 不同于传统特征蒸馏方法利用特征图间的绝对差值或均方误差计算蒸馏损失, 本文将 KL 散度直接作用于 Softmax 归一化后的局部特征上, 以约束特征的概率分布. 最后所有特征对之间计算出的损失数值相加, 作为最终的蒸馏损失. 局部特征蒸馏损失 L_{LFD} 的公式表述如公式 (6):

$$L_{\mathrm{LFD}}=\sum_{i=1}^{n}D_{\mathrm{KL}}(s_i|t_i) \tag{6}$$

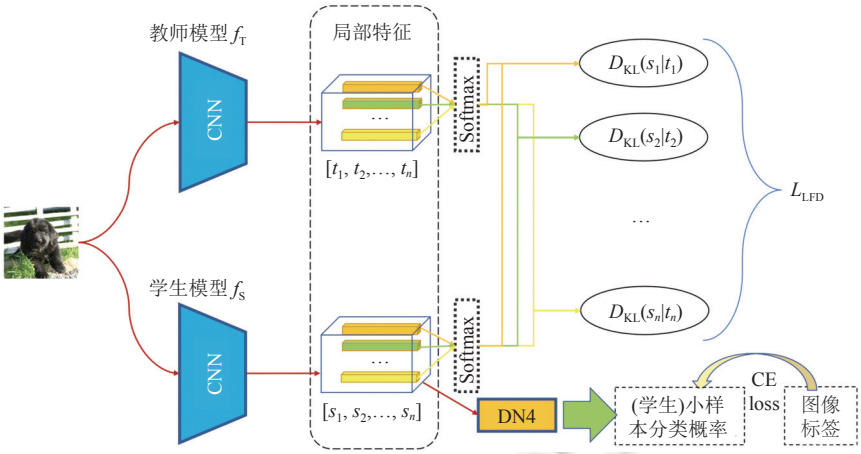


图 5 局部特征蒸馏 (LFD) 方法图示

如此, 局部特征蒸馏驱使学生模型的输出在每个局部位置上都一一对应地模拟教师模型的输出. 由于 DN4 分类器本身不包含可训练参数, 其度量分类的效果仅取决于特征提取网络的性能. 因此, 如果学生模型在图像样本上产生的特征分布趋近教师模型, 其小样本分类的精度就必然接近教师模型. 通过这种设计, 回避了传统蒸馏方法中蒸馏损失计算与类别本身的相关性, 转而追求学生模型在特征分布层次对教师的拟合, 更适用于小样本训练过程中产生的大量随机子任务. 而空间局部特征的使用, 保证了小样本任务在样本量受限的情况下也能获取大量的特征对, 增强了蒸馏作用的有效性.

3.3 基于辅助分类器的蒸馏方法

传统蒸馏方法应用于小样本学习任务时, 教师模型在预训练阶段得到的线性分类器在学生模型的小样本训练阶段 (即蒸馏阶段) 被完全弃用. 这一线性分类器产生的正是完整训练集类别空间上的固定概率分布. 那么, 是否有可能改变传统的直接在小样本任务分类概率上计算蒸馏损失的做法, 转而利用这种常规分类的概率分布提供额外监督信息. 基于这一想法, 本文设计了一种新的基于辅助分类器的蒸馏方法 (auxiliary classifier distillation, ACD).

辅助分类器蒸馏方法的基本流程如图 6 所示. 辅助分类器蒸馏方法中, 教师模型尾部不再套接 DN4 小样本度量分类器, 而是保留原有的 *GAP* 层和线性分类器. 每个小样本子任务中输入的一批图像数据也不再进行支持集和查询集的划分, 转而作为一个整体经特征提取网络及 *GAP* 层得到全局特征, 再输入给线性分类器, 得到训练集所有类别上的整体分类概率. 由于维度不同, 这种整体分类概率显然并不能直接用于指导学生模型的小样本分类训练. 对此, 本文给学生模型添加了一个额外的辅助线性分类器.

如此, 训练过程中学生模型的分类包含了两个彼此独立的分支: 第 1 分支中, 所有样本通过特征提取网络和 *GAP* 层, 得到全局特征, 输入给学生模型的辅助线性分类器, 得到训练集类别上的整体分类概率. 这一分类概率结果与教师模型的输出概率同维, 可用于计算蒸馏损失; 第 2 分支中, 输入样本经过特征提取网络 (不含 *GAP* 层) 得到局部特征, 进行支持集和查询集的划分后, 输入给小样本度量分类器, 得到小样本子任务上的分类概率. 小样本分类概率和真实标签之间计算交叉熵损失. 在测试阶段, 学生模型的辅助线性分类器即可弃用, 只使用度量分类器进行小样本任务测试. 辅助分类器蒸馏损失 L_{ACD} 如公式 (7) 所示, 其定义同公式 (1), 其中 s, t 分别表示学生模型和教师模型经线性分类器得到的分类概率, 长度等于整个训练集所含类别数.

$$L_{ACD} = D_{KL}(s|t) \tag{7}$$

如上设计中, 学生模型的辅助线性分类器虽然并未直接参与小样本分类任务, 但可以通过计算蒸馏损失及梯度回传对特征提取网络的参数进行更新, 增强学生模型特征提取的能力, 进而间接提升小样本学习的精度. 教师模型和学生模型都由常规线性分类器得到所有训练集类别上的固定分类概率, 并据此计算蒸馏损失, 避免了小样本随机任务中类别组合的不稳定性, 也为学生模型的蒸馏训练提供了一个相对恒定的更新目标.

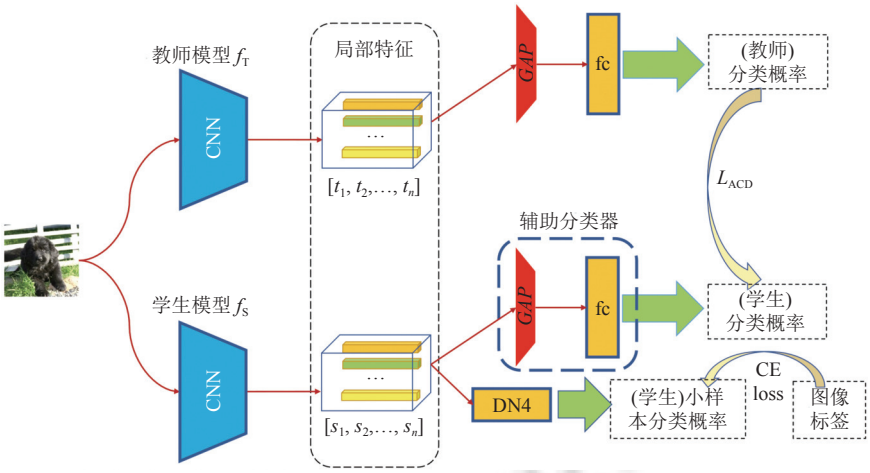


图 6 辅助分类器蒸馏方法 (ACD) 图示

4 实 验

4.1 数据集

本文的实验主要基于经典的小样本学习图像数据集 miniImageNet[22] 和 TieredImageNet[23]开展. miniImageNet 数据集是图像数据集 ImageNet[24]的一个子集, 共包含 100 个图像类别, 每类包含 600 张图像. 其中, 64 个类别用作训练集, 20 个类别用作训练后测试, 16 个类别用作训练过程中的验证. TieredImageNet 数据集也是 ImageNet 的子集, 其样本容量相对更大, 训练集-验证集-测试集分别包含 35 197 160 个图像类, 具体数据如表 1.

表 1 小样本学习实验数据集

数据集	训练集类别	验证集类别	测试集类别	总图像数
miniImageNet	64	16	20	60 000
TieredImageNet	351	97	160	779 165

实验中的图像加载遵循小样本学习任务中的通用设置: 输入图像采用的数据增广包括随机裁剪, 随机翻转和色彩抖动; 随机增广后图像的尺寸被统一缩放为 84×84 像素; 图像数值关于 ImageNet 数据集均值和方差的统计先验进行标准化. 小样本训练、验证和测试过程中, 每次都只抽取 N 个类别的图像样本构造 N-way K-shot 子任务. 训练集-验证集-测试集的类别空间均彼此互斥, 保证验证/测试时所用的类别都是模型在训练阶段未见的.

4.2 实验设置

教师模型的预训练采用传统的全局特征提取+线性分类方法, 在数据集的训练子集上进行. 预训练共 500 轮次, 初始学习率 1E–3, SGD 优化器, 学习率分别在第 75, 150, 300 轮次定点衰减 50%. 预训练过程中, 只使用分类概率与真实样本标签之间所得的分类交叉熵损失优化模型. 最终取训练过程中验证集上精度最高的轮次, 用于在后续实验中初始化教师模型权重.

为产生轻量化的学生模型, 需要对教师模型结构执行通道剪枝 (实验中剪枝率设定为 50%), 得到压缩后的学生模型网络结构 (Slim-ResNet12). 如表 2 所示, 除最初输入和最后输出通道外, 学生模型各层的 I/O 通道数相较教师模型缩减 50%, 整体可训练参数量压缩为教师模型的约 1/3. 剪枝后学生模型的参数全部重新随机初始化, 其中各卷积层应用 Kaiming normal 初始化[25], 各 BN 层缩放系数和偏移分别初始化为 1 和 0.

特征提取阶段, 对于统一尺寸的输入图像, 残差网络 ResNet12 和压缩后的 Slim-ResNet12 都为每张图像生成一组 5×5 个局部特征向量. 度量分类阶段, 学生模型在特征提取网络后套接 DN4 度量分类器完成小样本分类任

务, 分类器每个局部特征的近邻数设置为 3. DN4 分类器会根据支持集样本所含的局部特征, 预测每张查询集样本的小样本分类概率.

表 2 教师模型与学生模型网络结构对比

模型	网络结构	各block I/O通道数	参数量 (占比) (M)*	权重文件大小 (占比) (MB)
教师模型	ResNet12	3-64-160-320-640	12.65 (100%)	47.46 (100%)
学生模型	Slim-ResNet12	3-32-80-160-640	4.31 (≈34%)	15.63 (≈32%)

注: *教师模型参数量仅统计特征提取网络部分, 不包含其在预训练阶段用到的线性分类器. 小样本度量分类器不含可训练参数

学生模型的训练直接按照小样本“episodic training”方式从头进行, 利用传统分类交叉熵损失和蒸馏损失对模型进行联合优化. 所有实验中, 分类交叉熵损失都由小样本分类概率与真实标签计算得到. 而蒸馏损失根据不同的定义方式, 分为传统蒸馏, 局部特征蒸馏, 辅助分类器蒸馏 3 种方法, 分别进行实验.

(1) 传统蒸馏方法 (KD): 教师模型移除预训练阶段的线性分类器和 *GAP* 层, 套接 DN4 小样本度量分类器. 每个子任务的随机样本分别通过师生模型的特征提取网络和度量分类器, 得到当前子任务类别上的两组分类概率, 利用 KL 散度计算得到当前子任务的蒸馏损失.

(2) 局部特征蒸馏方法 (LFD): 教师模型移除线性分类器, 也不套接度量分类器, 只生成每批图像的局部特征作为中间结果, 与学生模型所得的同维局部特征按照公式 (6) 计算蒸馏损失.

(3) 辅助分类器蒸馏方法 (ACD): 教师模型保留预训练阶段的线性分类器, 学生模型除小样本度量分类器外, 另外附加一个常规线性分类器. 学生模型蒸馏训练阶段, 每个子任务抽取的随机样本作为一个整体, 分别通过师生模型的特征提取网络和线性分类器, 得到整个训练集类别空间上的同维分类概率, 根据公式 (7) 基于 KL 散度定义计算蒸馏损失.

如公式 (8) 所示, 所得蒸馏损失数值利用蒸馏损失权重 α 加权后, 与分类交叉熵损失加和, 作为一个整体对学生模型进行更新. 蒸馏损失权重的具体取值根据不同的蒸馏方法分别设置.

$$L = L_{\mathrm{CE}} + \alpha \times L_{\mathrm{KD}} \tag{8}$$

小样本学习任务的实验设置参考 Ye 等人的工作[26]: 学生模型都在训练集上进行 200 轮小样本训练, 每轮抽取 2 000 个随机子任务. 每个子任务包含 $N = 5$ 个类别, 每个类别包含 $K = 1$ 或 $K = 5$ 个支持集样本. 初始学习率 1E–3, SGD 优化器, 学习率每隔 20 轮衰减 50%. 每轮训练完成后, 在验证集上抽取 600 个随机子任务进行性能验证, 保存验证集精度最优的模型用作最终测试. 最终测试随机抽取多达 10 000 个随机子任务, 将所有子任务的精度均值作为最终结果, 减少随机因素对结果的干扰.

4.3 实验结果

4.3.1 无蒸馏实验

首先让学生模型在没有额外蒸馏辅助的情况下, 使用 DN4 度量分类器, 仅基于分类交叉熵损失从头进行小样本任务训练. 表 3 展示了在 miniImageNet 和 TieredImageNet 数据集上, 分别采用 5-way 1-shot 和 5-way 5-shot 小样本任务设置时的实验效果. 同时列出完整教师模型自身性能指标以供对比. 教师模型结果中每项包含两个数值: 第 1 项为预训练后直接进行小样本测试所得精度, 第 2 项是预训练后进一步小样本微调所得精度. 后文相同.

表 3 无蒸馏对照结果 (%)

模型	miniImageNet		TieredImageNet	
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
教师模型 (ResNet12)*	59.41/65.80	75.99/79.83	50.77/68.79	70.27/83.37
学生模型 (Slim-ResNet12)	56.15	74.67	56.46	75.72

注: *教师模型每项结果分别为“预训练后直接进行小样本测试”所得精度及“预训练后进一步小样本微调”所得精度

实验结果显示, 剪枝后的轻量化学生模型, 由于参数量大幅减少, 在没有额外蒸馏损失的情况下从头开始进行小样本分类训练, 最终小样本测试性能相较于完整教师模型有明显下降. 例如在 miniImageNet 上所得 5-way 1-shot

小样本任务测试精度仅为 56.15%. 不仅大幅低于教师模型经过小样本微调后的性能 (65.80%), 也低于教师模型预训练后直接进行小样本测试的性能 (59.41%).

4.3.2 小样本蒸馏实验

本节在 DN4 小样本度量分类的基础上添加第 4.2 节中介绍的各种蒸馏方法, 在不同数据集上分别进行实验. 结果如表 4 和表 5 所示. 其中蒸馏损失权重一栏为实验调试所得的每种蒸馏方法的最优权重设置. 如无特别说明, 表中各组蒸馏实验都按照公式 (8), 同时使用传统分类交叉熵损失和加权后的蒸馏损失, 对轻量化的学生模型进行联合优化.

表 4 miniImageNet 数据集小样本蒸馏结果

蒸馏方法	蒸馏损失权重	测试集精度 (%)	
		5-way 1-shot	5-way 5-shot
无	—	56.15	74.67
KD	0.05	60.31	77.93
LFD (ours)	0.5	63.84	79.49
ACD (ours)	2.0	**64.85**	**80.45**

表 5 TieredImageNet 数据集小样本蒸馏结果

蒸馏方法	蒸馏损失权重	测试集精度 (%)	
		5-way 1-shot	5-way 5-shot
无	—	56.46	75.72
KD	0.05	59.59	78.24
LFD (ours)	0.5	62.99	78.71
ACD (ours)	2.0	**63.38**	**79.31**

相较于无蒸馏的对照组, 各种蒸馏方法都对学生模型的小样本分类性能有明显的提升. 即使是直接基于小样本分类概率进行计算的传统蒸馏方法, 也将最终测试集精度提升了 3–4 个百分点.

相较于传统蒸馏方法, 本文设计的两类小样本蒸馏方法对最终精度的提升效果更为突出. 在 miniImageNet 数据集上, 局部特征蒸馏方法的最终测试精度达到 63.84% (+7.69%); 辅助分类器蒸馏方法最终精度更是高达 64.85% (+8.70%). 在 TieredImageNet 数据集上, 局部特征蒸馏方法最终测试集精度 62.99% (+6.53%), 辅助分类器蒸馏方法最终精度 63.38% (+6.92%), 相较传统蒸馏方法均有明显提升.

实验结果显示, 专门针对小样本学习任务设计的蒸馏方法相较于传统蒸馏方法有显著的优势, 证明本文新定义的蒸馏方法更适合小样本学习任务. 通过直接从特征层次进行约束, 或者通过额外的固定分类概率进行约束, 有效规避了小样本学习中每个子任务类别数较少且随机组合的限制, 为学生模型的训练提供一个相对稳定的更新目标, 从而更充分地发掘学生模型的潜在性能.

4.3.3 对比其他小样本学习方法

表 6 展示了本文方法与部分使用 ResNet12 作为特征提取网络的其他小样本学习方法的性能对比. 由于本文主要基于轻量化设置开展实验, 结果直接与使用完整 ResNet12 的其他方法比较存在一定不公平. 因此, 本文复现并补充了 ProtoNet, DN4 两种无参小样本度量分类方法使用 Slim-ResNet12 轻量化模型, 从头进行小样本训练所得的结果. 完整 ResNet12 和轻量化 Slim-ResNet12 各自的最优结果加粗标识.

表 6 本文方法与其他小样本学习方法性能比较 (%)

模型	小样本学习方法	知识蒸馏	miniImageNet		TieredImageNet	
			5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
ResNet12 (12.65M)	Baseline++[27]	否	56.39	76.18	65.54	83.46
	ProtoNet[16]	否	57.10	74.20	62.93	83.30
	DN4[21]	否	59.14	75.26	64.41	82.59
	RelationNet[28]	否	55.22	69.25	56.86	74.66
	CAN[29]	否	62.68	78.36	70.46	84.50
	RENet[30]	否	64.81	79.90	70.14	82.70
	FEAT[26]	否	**66.78**	**82.05**	**70.80**	**84.79**
Slim-ResNet12 (4.31M)	ProtoNet[16]	否	48.49	73.31	40.31	72.19
	DN4[21]	否	56.15	74.67	56.46	75.72
	DN4+LFD (ours)	是	63.84	79.49	62.99	78.71
	DN4+ACD (ours)	是	**64.85**	**80.45**	**63.38**	**79.31**

由表 6 结果可见, 在基于 Slim-ResNet12 的轻量化设置下, 本文提出的两种小样本蒸馏方法的结果, 显著优于未使用蒸馏的两组对照. 但与使用完整 ResNet12 的方法相比, 轻量化蒸馏的结果仍有一定差距.

4.3.4 小样本学习模型轻量化效果

表 7 展示了完整 ResNet12 和蒸馏所得的轻量化 Slim-ResNet12 各自的参数规模和运行指标. 测试任务利用 DN4 小样本分类器在 miniImageNet 测试集上 (包含 10 000 个随机任务) 基于 5-way 1-shot 设置进行. 测试设备为单张 32 GB V100 显卡. 其中推理时间为 3 次测试所得的平均值. 对比可见, 轻量化蒸馏对于小样本学习模型起到了明显的压缩和加速的效果, 有利于模型在计算资源受限情况下的部署应用.

表 7 轻量化蒸馏压缩及加速效果

模型	参数量 (M)/占比 (%)	推理时显存占用 (MB)/占比 (%)	测试集推理用时 (s)/占比 (%)
ResNet12	12.65/100	3 607/100	1 055/100
Slim-ResNet12	4.31/约34	2 531/约70	885/约84

4.4 补充分析

4.4.1 蒸馏损失权重

蒸馏损失权重用于在学生模型训练过程中平衡蒸馏作用的重要程度. 权重设置过小, 蒸馏无法起到应有的效果; 设置过大, 又有可能干扰小样本分类任务本身. 对本文提出的两种蒸馏方法, 在 miniImageNet 数据集上按照 5-way 1-shot 设置进行小样本蒸馏实验. 分别调试蒸馏损失权重, 记录最优验证集精度和最终的测试集精度. 结果如表 8 所示.

表 8 蒸馏损失权重对结果的影响 (%)

蒸馏方法	蒸馏损失权重	验证集精度 5-way 1-shot	测试集精度 5-way 1-shot
LFD	0.01	58.32	56.95
	0.05	65.79	63.21
	0.1	65.95	63.49
	0.5	66.22	**63.84**
	1.0	66.38	63.50
	2.0	66.42	63.75
ACD	0.1	66.74	63.56
	0.5	66.13	63.70
	1.0	66.51	64.65
	2.0	65.84	**64.85**
	5.0	*	*

注: *表示蒸馏损失数值过大, 造成梯度爆炸, 训练无效

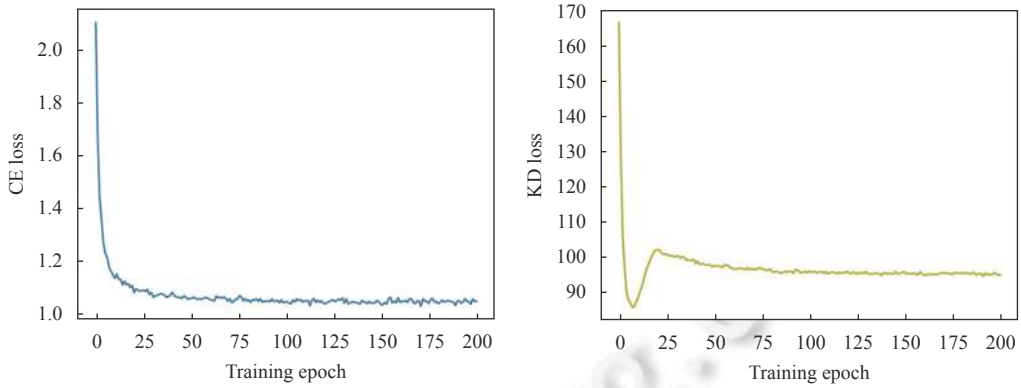
以局部特征蒸馏方法在蒸馏损失权重 0.5 下所得的实验记录为例, 统计学生模型训练过程中分类交叉熵损失和蒸馏损失数值 (加权后) 的变化情况, 结果如后文图 7 所示. 分类损失数值从开始一直平稳下降, 训练半程后基本保持稳定. 而蒸馏损失在训练的最初十几轮有一定浮动, 之后也逐渐平稳下降. 损失数值的变化表明随着蒸馏训练的进行, 学生模型小样本分类能力稳步提升的同时, 其提取图像特征的能力也逐渐趋近于教师模型.

从实验记录中还可以看出, 在各组蒸馏实验的运行过程中, 蒸馏损失 (加权后) 在数值上远大于分类交叉熵 (CE) 损失, 其绝对大小甚至超出 1–2 个数量级. 说明来自教师模型的“软标签”作为一种辅助监督信息, 有时能比真实标签对学生模型起到更强的指导作用. 但这并不意味着常规的小样本分类交叉熵损失是无用的. 后文表 9 显示了仅用局部特征蒸馏损失对模型进行更新的效果, 相较同时使用两种损失对模型进行联合优化的结果, 测试集精度降低了约 0.8%.

4.4.2 多种损失联合使用

本文也尝试了联合使用多种蒸馏损失的实验, 并对蒸馏损失的相对权重进行了调试. 在两种蒸馏方法各自最

优权重的基础上, 分别向下调整, 设计了 5 组实验. 表 10 显示在 miniImageNet 数据集上, 5-way 1-shot 任务设置下, 联合使用两种蒸馏损失对学生模型最终测试集精度的影响.



(a) CE loss (b) KD loss

图 7 学生模型训练过程中交叉熵损失及蒸馏损失数值变化

表 9 分类交叉熵损失对结果的影响 (%)

损失方案	蒸馏损失权重	5-way 1-shot	
		验证集精度	测试集精度
LFD + CE	0.5	66.22	**63.84**
仅LFD	0.5	65.67	63.02

表 10 多种损失联合使用 LFD+ACD 对结果的影响 (%)

蒸馏损失权重	5-way 1-shot	
	验证集精度	测试集精度
0.05+0.05	65.97	63.54
0.1+0.1	65.77	63.06
0.1+1.0	**66.88**	**64.06**
0.5+1.0	66.45	64.04
0.5+2.0	66.12	63.72

根据实验结果, 联合使用多种蒸馏损失并不能带来进一步的性能提升. 其最终测试集精度并未超过单独使用某一蒸馏损失方法时的最高值 (64.06% < 64.85%). 造成这一现象的原因可能是两种蒸馏方法向轻量化学生模型传递的潜在知识存在一定差异. 具体而言, 局部特征蒸馏方法强调中间层特征分布层面的接近, 而辅助分类器蒸馏方法更侧重分类概率层面的拟合. 两种蒸馏方法对模型产生的权重更新方向不完全一致, 因此联合使用并不能起到性能简单叠加的效果.

4.4.3 局部特征蒸馏实现方案

除了前文介绍的直接基于图像局部特征数值进行约束的方法, 本文也尝试借鉴基于关系的蒸馏方法[9,10], 进一步探索局部特征蒸馏的实现方式. 现有的基于关系蒸馏的方法, 大多利用同一批次样本间的相互关系作为蒸馏凭据. 然而在小样本学习任务中, 这种做法同样面临单个任务所含样本数较少的局限性. 对此, 本文尝试把样本间的相互关系改进为单个样本内部局部特征间的相互关系, 通过特征间关系约束实现局部特征蒸馏. 该方案下, 蒸馏损失由师生模型对同一样本各自所得的同维特征图之间计算均方误差得到.

表 11 展示了两种局部特征蒸馏实现方案的对比. 实验在 miniImageNet 数据集上进行, 并分别记录两种方案在各自最优蒸馏损失权重下得到的结果.

对比结果显示, 基于图像内部的局部特征间关系进行蒸馏约束, 对学生模型的性能也有明显的提升, 且最优结果仅略低于基于特征数值约束的方案 (5-way 1-shot 任务上相差 0.27%). 可见, 对于利用图像局部特征进行小样本分类的模型, 图像内部特征间的相对关系也蕴含着重要的潜在知识, 可以通过蒸馏手段辅助轻量级模型的高效学习. 相较于样本间的关系, 这种局部特征间关系包含的相互信息更加丰富, 也因此更适用于每个子任务只包含少量样本的小样本学习任务.

表 11 局部特征蒸馏实现方案对结果的影响 (%)

蒸馏方法	蒸馏损失权重	测试集精度	
		5-way 1-shot	5-way 5-shot
无	—	56.15	74.67
LFD (基于特征数值约束)	0.5	**63.84**	**79.49**
LFD (基于特征间关系约束)	0.1	63.57	79.47

4.4.4 自蒸馏效果

本文主要基于小样本学习模型轻量化任务展开实验, 但新提出的小样本蒸馏方法其实并不局限于轻量化任务, 在非轻量化背景下应该也是有效的. 对此, 本节补充了小样本学习模型在自蒸馏设置 (即学生模型采用和教师模型相同的网络结构和参数规模) 下, 使用局部特征蒸馏方案在 miniImageNet 数据集上的实验.

首先对完整的 ResNet12 网络进行预训练, 所得网络权重被同时用于初始化教师和学生模型. 学生模型在预训练的基础上, 再按照小样本“episodic training”方式进行蒸馏训练, 利用分类交叉熵损失和局部特征蒸馏损失进行联合优化. 自蒸馏实验结果如表 12 所示.

表 12 自蒸馏设置下 miniImageNet 数据集结果 (%)

模型	蒸馏方法	测试集精度	
		5-way 1-shot	5-way 5-shot
教师模型 (ResNet12)*	无	59.41/65.80	75.99/79.83
学生模型 (ResNet12)	LFD	**66.80**	**81.20**

注: *教师模型每项结果分别为“预训练后直接进行小样本测试”所得精度及“预训练后进一步小样本微调”所得精度

在自蒸馏设置下, 学生模型最终的小样本测试集精度分别达到了 66.80% 和 81.20%, 这一结果甚至超过教师模型自身微调后所得的精度 (65.80% 和 79.83%). 实验结果显示, 局部特征蒸馏方法能在没有引入额外训练数据的情况下, 对深度网络本身的性能做更深入的挖掘. 进一步证明了该蒸馏方法可以作为小样本学习模型训练中提升精度的一种辅助手段. 达到这一指标的一个重要原因是学生模型从预训练权重初始化, 完整继承了预训练阶段学习到的特征提取性能. 相比之下, 前文的轻量化蒸馏实验中, 学生模型都是随机初始化后从头开始训练的.

5 总 结

小样本学习作为深度学习领域的重要分支, 具有广阔的应用前景, 但专门针对小样本学习的模型轻量化研究目前仍相当有限. 本文面向小样本学习任务, 引入知识蒸馏策略实现小样本学习模型的轻量化. 针对小样本学习训练过程中包含大量随机子任务, 且子任务类别随机组合的特点, 本文设计了局部特征蒸馏和辅助分类器蒸馏这两类新的小样本蒸馏方法. 通过小样本任务数据集上的蒸馏实验, 证明了本文方法能够大幅提升小样本学习轻量化模型的潜在性能, 且相较于传统知识蒸馏方法具有显著的优势.

希望本文的工作能对将来更多围绕小样本学习模型轻量化的研究提供参考和启发.

References:

[1] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[2] Zhang LF, Song JB, Gao AN, Chen JW, Bao CL, Ma KS. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3712–3721. [doi: 10.1109/ICCV.2019.00381]

[3] Yun S, Park J, Lee K, Shin J. Regularizing class-wise predictions via self-knowledge distillation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 13873–13882. [doi: 10.1109/CVPR42600.2020.01389]

[4] Phuong M, Lampert C. Distillation-based training for multi-exit architectures. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1355–1364. [doi: 10.1109/ICCV.2019.00144]

[5] Li TH, Li JG, Liu Z, Zhang CS. Few sample knowledge distillation for efficient network compression. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 14627–14635. [doi: 10.1109/CVPR42600.2020.01465]

[6] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[7] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego, 2015.

[8] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[9] Tung F, Mori G. Similarity-preserving knowledge distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1365–1374. [doi: 10.1109/ICCV.2019.00145]

[10] Park W, Kim D, Lu Y, Cho M. Relational knowledge distillation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3967–3976. [doi: 10.1109/CVPR.2019.00409]

[11] Liu Y, Zhang W, Wang J. Learning from a lightweight teacher for efficient knowledge distillation. arXiv:2005.09163, 2020.

[12] Hou YN, Ma Z, Liu CX, Loy CC. Learning lightweight lane detection CNNs by self attention distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1013–1021. [doi: 10.1109/ICCV.2019.00110]

[13] Han S, Mao HZ, Dally WJ. Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding. In: Proc. of the 4th Int'l Conf. on Learning Representations. San Juan, 2016.

[14] Jankowski N, Duch W, Grąbczewski K. Meta-learning in Computational Intelligence. Berlin: Springer, 2011. [doi: 10.1007/978-3-642-20980-2]

[15] Wang YX, Hebert M. Learning to learn: Model regression networks for easy small sample learning. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 616–634. [doi: 10.1007/978-3-319-46466-4_37]

[16] Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4080–4090.

[17] Li D, Zhang JS, Yang YX, Liu C, Song YZ, Hospedales T. Episodic training for domain generalization. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1446–1455. [doi: 10.1109/ICCV.2019.00153]

[18] Tian YL, Wang Y, Krishnan D, Tenenbaum JB, Isola P. Rethinking few-shot image classification: A good embedding is all you need? In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 266–282. [doi: 10.1007/978-3-030-58568-6_16]

[19] Rajasegaran J, Khan S, Hayat M, Khan FS, Shah M. Self-supervised knowledge distillation for few-shot learning. In: Proc. of the 32nd British Machine Vision Conf. BMVA Press, 2021. 179.

[20] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[21] Li WB, Wang L, Xu JL, Huo J, Gao Y, Luo JB. Revisiting local descriptor based image-to-class measure for few-shot learning. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7260–7268. [doi: 10.1109/CVPR.2019.00743]

[22] Malalur P, Jaakkola T. Alignment based matching networks for one-shot classification and open-set recognition. arXiv:1903.06538, 2019.

[23] Ren MY, Triantafillou E, Ravi S, Snell J, Swersky K, Tenenbaum JB, Larochelle H, Zemel RS. Meta-learning for semi-supervised few-shot classification. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[24] Deng J, Dong W, Socher R, Li LJ, Li K, Li FF. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[25] He KM, Zhang XY, Ren SQ, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 1026–1034. [doi: 10.1109/ICCV.2015.123]

[26] Ye HJ, Hu HX, Zhan DC, Sha F. Few-shot learning via embedding adaptation with set-to-set functions. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8805–8814. [doi: 10.1109/CVPR42600.2020.00883]

[27] Chen WY, Liu YC, Kira Z, Wang YCF, Huang JB. A closer look at few-shot classification. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[28] Sung F, Yang YX, Zhang L, Xiang T, Torr PHS, Hospedales TM. Learning to compare: Relation network for few-shot learning. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1199–1208. [doi: 10.1109/CVPR.2018.00131]

[29] Hou RB, Chang H, Ma BP, Shan SG, Chen XL. Cross attention network for few-shot classification. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver, 2019. 360.

[30] Kang D, Kwon H, Min JH, Cho M. Relational embedding for few-shot classification. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 8802–8813. [doi: 10.1109/ICCV48922.2021.00870]

陈嘉言(1997－), 男, 硕士, 主要研究领域为计算机视觉, 小样本学习.

霍静(1989－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为机器学习, 计算机视觉, 图像生成, 人脸识别.

任东东(1993－), 男, 博士生, 主要研究领域为计算机视觉, 软硬件协同.

高阳(1971－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为强化学习, 多智能体系统.

李文斌(1991－), 男, 博士, 副研究员, CCF 专业会员, 主要研究领域为机器学习, 计算机视觉, 小样本学习, 自监督学习, 持续学习.