控制与决策
Control and Decision

基于动态混合注意力的自知识蒸馏

唐媛, 陈莹

引用本文:
唐媛, 陈莹. 基于动态混合注意力的自知识蒸馏[J]. 控制与决策, 2024, 39(12): 4099-4108.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.0036

基于动态混合注意力的自知识蒸馏

唐 媛, 陈 莹†

(江南大学 轻工过程先进控制教育部重点实验室，江苏 无锡 214122)

摘 要: 自知识蒸馏降低了对预训练教师网络的依赖,但是其注意力机制只关注图像的主体部分,一方面忽略了携带有颜色、纹理信息的背景知识,另一方面空间注意力的错误聚焦情况可能导致主体信息遗漏. 鉴于此,提出一种基于动态混合注意力的自知识蒸馏方法,合理挖掘图像的前背景知识,提高分类精度. 首先,设计一个掩膜分割模块,利用自教师网络建立注意力掩膜并分割出背景特征与主体特征,进而提取背景知识和遗漏的主体信息;然后,提出基于动态注意力分配策略的知识提取模块,通过引入基于预测概率分布的参数动态调整背景注意力和主体注意力的损失占比,引导前背景知识相互协作,逐步优化分类器网络对图像的关注,提高分类器网络性能. 实验结果表明: 所提出方法使用 ResNet 18 网络和 WRN-16-2 网络在 CIFAR 100 数据集上的准确率分别提升了 2.15 % 和 1.54 %;对于细粒度视觉识别任务,使用 ResNet 18 网络在 CUB 200 数据集和 MIT 67 数据集上的准确率分别提高了 3.51 % 和 1.05 %,其性能优于现有方法.

关键词: 深度学习；模型压缩；知识蒸馏；图像分类；注意力机制；背景知识

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0036

引用格式: 唐媛,陈莹. 基于动态混合注意力的自知识蒸馏[J]. 控制与决策, 2024, 39(12): 4099-4108.

Self-knowledge distillation based on dynamic mixed attention

TANG Yuan, CHEN Ying†

(Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education，Jiangnan University，Wuxi 214122，China)

Abstract: Self-knowledge distillation reduces the necessity of training a large teacher network, whose attention mechanism only focuses on the foreground of the image. It ignores the background knowledge with color and texture information, furthermore may lead to the omission of the foreground information due to the wrong focus of spatial attention. To address the problem, a self-knowledge distillation method based on dynamic mixed attention is proposed, which reasonably exploits both foreground and background information in images and therefore improves the classification accuracy. A mask segmentation module is designed to segment the feature map of background and foreground, which are used to extract the ignored background knowledge and the missing foreground information respectively. Moreover, a knowledge extraction module based on dynamic attention distribution strategy is proposed, which dynamically adjusts the loss ratio of background attention and foreground attention by introducing a parameter based on predictive probability distribution. The strategy guides the cooperation between foreground and background, which leads to more accurate attention map and improves the performance of a classifier network. Experiments show that the proposed method using ResNet 18 and WRN-16-2 improves the accuracy on CIFAR 100 by 2.15 % and 1.54 % respectively. For fine-grained visual recognition tasks, the accuracy on CUB 200 dataset and MIT 67 dataset is improved by 3.51 % and 1.05 % respectively, which makes its performance superior to the state-of-the-arts.

Keywords: deep learning；model compression；knowledge distillation；image classification；attention mechanism；background knowledge

0 引 言

随着深度学习的不断发展,更大更深的模型表现出更好的性能,但其对计算资源和存储容量的高需求迫使人们探索减少参数量和压缩模型的方法. 知识蒸馏[1]旨在通过令小体积的学生模型学习大体积教师模型的知识并逼近教师模型输出,在图像分类[2]、

收稿日期: 2024-01-08；**录用日期:** 2024-04-24.
基金项目: 国家自然科学基金项目(62173160).
责任编委: 张文安.
†通讯作者. E-mail: chenying@jiangnan.edu.cn.

目标检测[3]、异常检测[4]等计算机视觉任务中取得了优异的表现. 传统的离线知识蒸馏方法依赖于笨重的教师模型,需要花费大量时间和计算资源来预先训练教师模型. 此外,对教师模型的依赖也带来了一个隐患,即不同结构的小体积学生模型难以匹配具有复杂表达空间的教师模型,降低了蒸馏效果. 因此,人们开始研究自知识蒸馏方法,其逐步训练学生模型来提取自身的知识,无需预训练的教师模型.

现有自知识蒸馏方法着眼于主体部分,在最终分类时以最后一层特征图作为输入进行预测,而在最后一层特征图中,背景部分的特征激活值相对于主体部分极小,因此,背景信息在预测过程中常常是被忽略的. 但是事实上,背景信息中包含了空间位置、颜色、纹理等信息,如绿色的树林、蓝色的海洋、水波的纹理等背景信息均可辅助模型学习,学习背景知识也能够增加蒸馏知识的多样性. 此外,在一些注意力划分不准确的情况下,背景中可能包含丢失的主体信息.

为了解决上述问题,本文提出基于动态混合注意力的自知识蒸馏方法(self-knowledge distillation based on dynamic mixed attention, MA-SKD),在训练过程中利用自教师网络的空间注意力将最后一层特征图掩膜分割为主体特征图和背景特征图,并设计具有动态注意力分配策略的知识提取模块,根据携带信息量不同的样本改变知识提取的重心,通过动态提取被遗漏的主体知识和被忽略的背景知识,引导前背景知识相互协作,逐步优化网络对图像的关注,进而提升分类器网络性能.

1 相关工作

1.1 知识蒸馏

针对大规模神经网络的实际部署和应用,模型压缩方法成为近期的研究热点. 现有的神经网络压缩方法主要包括网络剪枝[5]、低秩分解[6]、数据量化[7]、知识蒸馏[1]和紧凑网络设计[8],其中知识蒸馏凭借直观、高效的知识迁移思想以及优越的模型压缩性能受到研究者的广泛关注. 知识蒸馏框架由Hinton等[9]首次提出,其核心思想是从预先训练好的高性能大规模模型(教师模型)中转移知识给小体积模型(学生模型),令学生模型产生与教师模型相近的输出,从而提高学生模型的性能. 这里的知识包括隐藏层特征、最终层的分类概率等.

根据蒸馏方案的不同,知识蒸馏可分为离线蒸馏、在线蒸馏和自知识蒸馏[1]. 离线蒸馏即传统的“学生-教师”架构的知识蒸馏. 潘瑞东等[10]提出了基于BERT的多任务多标签文本分类模型,以教师模型评价为蒸馏知识指导学生训练;程旗等[11]利用教师模型的RPN输入特征知识提升了学生模型性能;ER-KD[12]利用教师模型的熵对单个样本的损失进行重新加权,缓解了学生模型过度自信的预测;CoMD[13]建立了教师指导学生和学生输出反馈给教师的双向反馈回路,解决了基于指令调优导致单向性知识转移问题. 在线蒸馏旨在令多个学生网络互相协作,互相优化,无需教师模型的指导. Li等[14]通过构建相同架构的公共学生集和一个领导学生,分别得到融合特征图和解码的浅层特征图,令信息从较深的网络层流向较浅的网络层,增加了训练过程的信息丰富性;ORCKD[15]每轮将多个子网络划分为教师群体和学生群体,使用强化教学、私教和小组教学多种教学模式指导学生训练.

1.2 自知识蒸馏

自知识蒸馏是在没有教师模型的情况下,逐步训练学生模型来提取自身的知识[1]. 自知识蒸馏大致可分为两类:基于数据增强的方法和基于辅助网络的方法.

基于数据增强的方法侧重于在同一样本的不同版本间传递知识. Xu等[16]提出了学习相同样本的不同增强版本间的一致性特征,以此作为知识引导学生学习;CS-KD[17]提出了类正则化项,强制同一类的不同样本产生一致预测;MixSKD[18]提取了原始图像及其混合图像的随机对间的特征图和概率分布,并聚合多级特征图作为软标签. 此类方法利用数据增强来扩充数据量,学习不同增强版本间的不变性,但是不同版本的失真实例易丢失它们的局部信息.

基于辅助网络的方法通过引入辅助网络,诱导分类器网络的辅助分支产生相似的输出,或令学生网络的中间特征图对齐辅助网络在训练中提取的细化特征图. BYOT[19]通过引入辅助分类器网络得到了隐藏层的分类概率,并结合最终层的logit来指导学生网络的训练;FRSKD[20]利用辅助自教师网络获得各层的精细化特征图,作为软标签指导学生模型学习;PR-SKD[21]通过金字塔结构细化特征,在保持浅层特征信息的基础上增强了深层特征的能力. 此类方法引入与学生网络相似的辅助网络得到中间层的概率分布或精细化知识,但是本质上学习的还是每层对应主体部分的知识.

此外,一些自知识蒸馏方法改进了蒸馏过程. Tf-KD[22]预先训练学生模型,并利用其输出作为辅助教师的软标签来训练同结构学生模型. Zipf's LS[23]使用即时预测来生成符合Zipf分布的监督信息,并生成

实例级非均匀软标签来引导网络; AI-KD [24] 从历史预测的概率分布中逐步提取知识, 使用对抗学习转移概率分布知识; DRG+DSR [25] 通过融合不同级别的信息构建了一个信息更丰富的虚拟教师, 设计了一种正则化提取方法以确保所有数据的输出具有一致性; USKD [26] 将一般KD损失分解和重组为标准化KD损失, 并为目标类和非目标类定制软标签.

2　基于动态混合注意力的自知识蒸馏方法

2.1　基础知识

传统的知识蒸馏中学生向教师网络的软标签和真实标签学习. 给定一个由 θ_s 参数化的学生网络 f_s 和由 θ_t 参数化的自教师网络 f_t, 知识蒸馏的损失表达式如下所示:

$$L_{\mathrm{KD}} = n \cdot L_{\mathrm{CE}}(f_s(x), y) + (1-n) \cdot L_{\mathrm{KL}}(x, y; \theta_t, \theta_s, T). \quad (1)$$

其中: L_{CE} 为交叉熵损失, L_{KL} 为KL散度损失, x 是真实标签为 y 的输入样本, $T>0$ 为温度参数, n 为超参数. 为了减少对预训练教师网络的依赖, 一些自知识蒸馏方法引入辅助网络来提取自身的知识. FRSKD为本文的基准网络, 它引入基于BiFPN [27] 网络的辅助自教师网络, 通过特征融合得到网络中间各层的精细化特征图, 以此指导分类器网络(学生网络)学习. 令 $F_t^i(i \in [1,2,\ldots,N])$ 和 $F_s^i(i \in [1,2,\ldots,N])$ 分别为自教师网络和分类器网络的 N 层特征图. FRSKD方法的优化目标如下所示:

$$L_{\mathrm{FRSKD}}(x, y; \theta_t, \theta_s; T) = L_{\mathrm{CE}}(f_s(x), y) + L_{\mathrm{CE}}(f_t(x), y) + \alpha \cdot L_{\mathrm{KL}}(x, y; \theta_t, \theta_s, T) + \beta \cdot L_F(F_t, F_s; \theta_t, \theta_s). \quad (2)$$

这里: $f_s(x)$ 和 $f_t(x)$ 分别为分类器网络和自教师网络产生的预测向量, L_F 为基于注意力转移的特征蒸馏损失, α 和 β 为超参数[20].

虽然FRSKD方法引入自教师网络得到了中间层的精细化特征图和最终的软标签, 但是考虑到两者网络结构和参数量的差距, 优化已有的特征图和软标签知识的提升效果有限. FRSKD方法与大多数知识蒸馏方法一样致力于关注图像的主体部分, 忽略了携带丰富的颜色、纹理信息的背景部分. 为了挖掘训练过程中的背景知识, 引导学生在学习主体部分的基础上利用额外的背景知识辅助学习, MA-SKD方法利用注意力掩膜分离得到背景部分和主体部分的特征激活值, 增加蒸馏知识的多样性, 提高分类器网络的预测精度.

2.2　网络整体框架

为了尽可能地挖掘背景信息, 提取训练过程中丢失的主体信息, 增加蒸馏知识的多样性从而辅助学生学习, 提出了基于动态混合注意力的背景知识提取算法MA-SKD. 该方法利用自教师网络的注意力对特

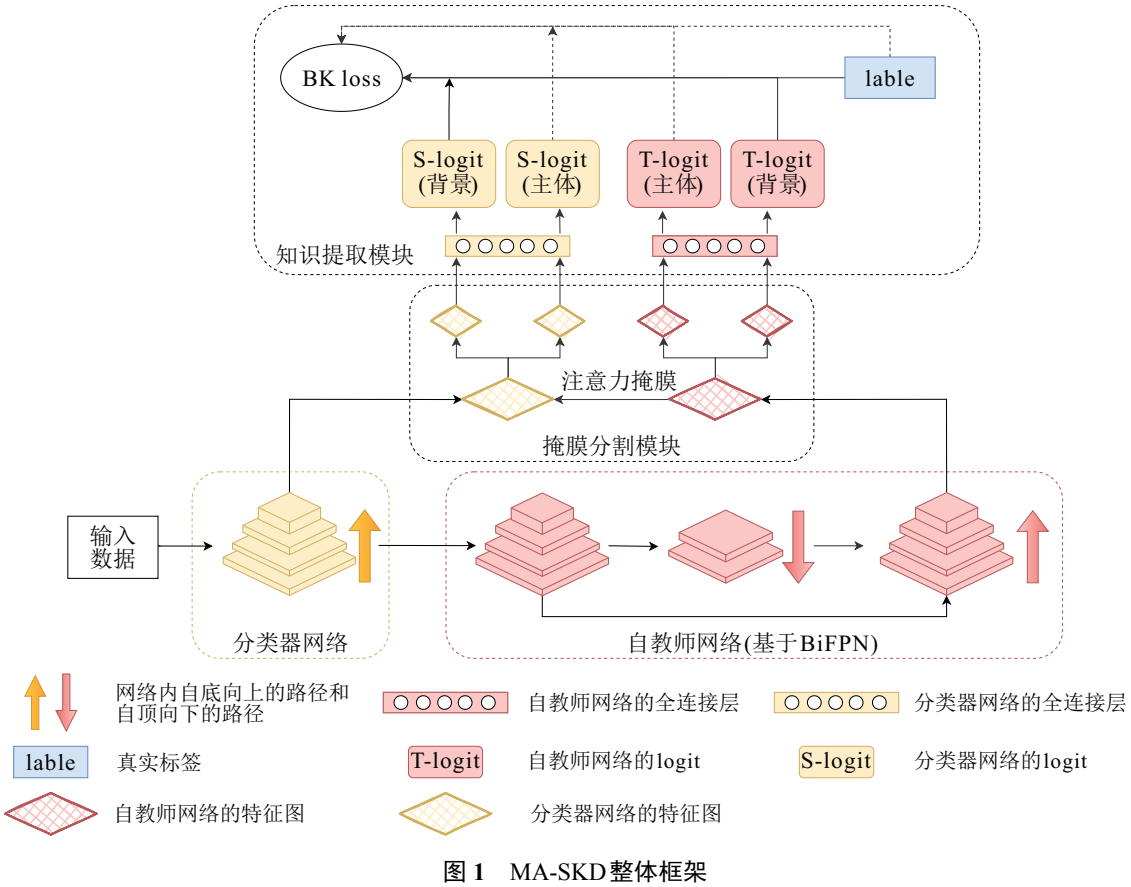


图 1　MA-SKD整体框架

征图进行掩膜分割,并蒸馏得到被忽略的背景知识和丢失的主体知识. 图1为MA-SKD的网络框架,其由4个模块组成:自教师网络、分类器网络、掩膜分割模块和知识提取模块.

分类器网络主要有4层,其以自底向上的路径提取输入样本的特征,经平均池化和全连接层得到S-logit. 分类器网络学习每层的精细化特征图并将自教师网络的T-logit作为软标签.

自教师网络采用来自PANet[28]和BiFPN[27]自顶向下的路径和自底向上的路径,每层输入来自分类器网络相应层提供的原始特征,与分类器网络同步进行训练. 自顶向下的路径用于聚合空间信息,聚合后的特征在自底向上的路径中被进一步提取,最终为分类器网络提供精细化的特征图.

掩膜分割模块以自教师网络和分类器网络的最后一层特征图为输入,以自教师网络的注意力为依据掩膜分割得到主体特征图和背景特征图. 知识提取模块以掩膜后的特征图作为输入,以概率分布作为输出,将背景知识和被遗漏的主体知识从自教师网络转移至分类器网络,并使用动态的注意力分配策略,根据训练进程调整对背景知识的关注度,针对携带信息量不同的训练样本,调整对主体和背景的关注重心.

2.3 掩膜分割模块

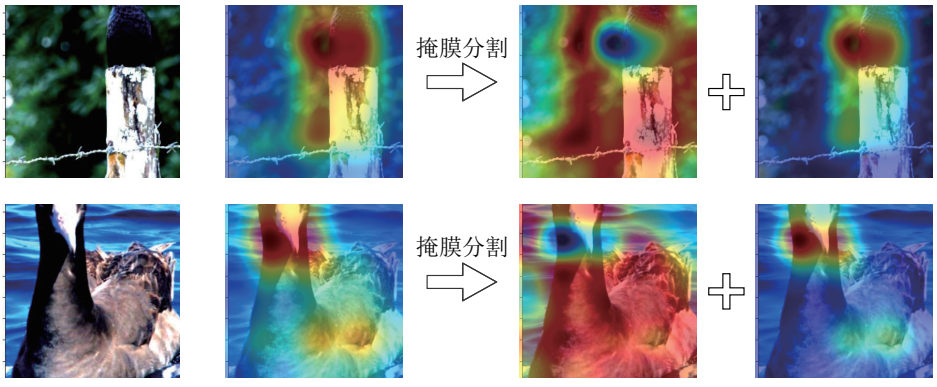
在传统的知识蒸馏中,以分类器网络的最后一层特征图为唯一输入,而与主体特征激活值相比数值较小的背景特征激活值在预测过程中被忽略. 因此,设计一个掩膜分割模块对特征图进行处理,分离主体特征激活值和背景特征激活值. 为了精确地对特征图进行分割,以精度较高的自教师网络的最后一层空间注意力为依据,分别对分类器网络和自教师网络的最后一层特征图进行掩膜分割,得到强调主体注意力的特征图和强调背景注意力的特征图.

主体特征图排除低数值背景部分的干扰,专注于主体部分轮廓,提取主体知识为分类任务作为补充;背景特征图去除高数值主体部分的干扰,提取背景知识作为额外的蒸馏知识,辅助分类器网络学习,使得分类器网络进一步逼近自教师网络的输出,从而产生更精准的注意力. 令最后一层特征图$t\in R^{C\times H\times W}$,使用自教师网络的空间注意力的掩膜分割过程为

$$M_b=\left(1-\frac{\sum_{i=1}^{C}t_i}{\max\left(\sum_{i=1}^{C}t_i\right)}\right)\odot t,$$

$$M_m=\left(\frac{\sum_{i=1}^{C}t_i}{\max\left(\sum_{i=1}^{C}t_i\right)}\right)\odot t. \tag{3}$$

其中:$\odot$为对应位置的矩阵点乘操作;C为通道数;$\sum_{i=1}^{C}t_i$为按照通道压缩求得对应的注意力图;M_b和M_m分别为掩膜分割得到的背景特征图和主体特征图,将其按照通道分别压缩为背景注意力图和主体注意力图进行可视化,最终的掩膜分割过程注意力可视化如图2所示.

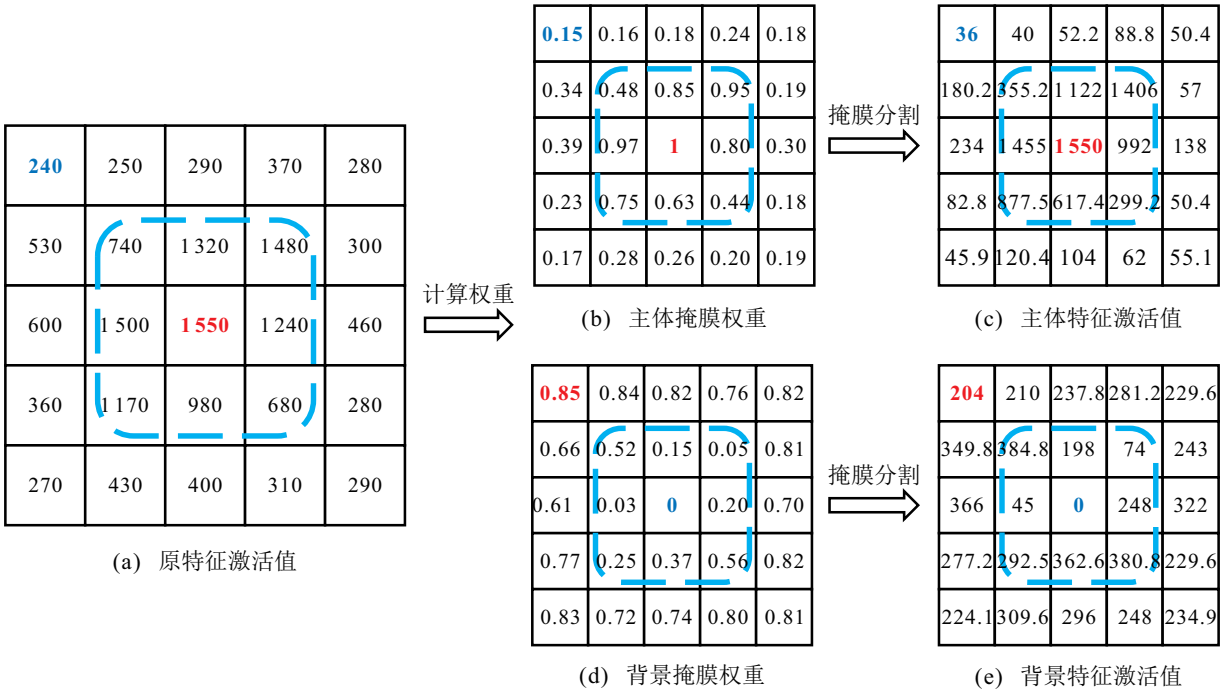


(a) 原图 (b) 原注意力 (c) 纯背景注意力 (d) 纯主体注意力

图2 掩膜分割过程注意力可视化

结合式(3)和图2分析可知:掩膜分割过程是将自教师网络的空间注意力归一化后在每个像素位置得到0～1间的主体权重值,对应位置的背景权重值通过1-主体权重值计算得到;利用背景权重和主体权重与原特征图激活值分别相乘得到对应的纯背景特征激活值和纯主体特征激活值,实现分割主体和背景特征图的目的.

为了进一步阐述掩膜分割过程,记原特征激活值为通道压缩后的特征激活值,由式(3)得到掩膜分割过程中特征激活值变化如图3所示. 其中:背景部分最小值位置用蓝色标注,主体部分最大值位置用红色标注,两者数据取自局部实验数据.



(a) 原特征激活值　(b) 主体掩膜权重　(c) 主体特征激活值　(d) 背景掩膜权重　(e) 背景特征激活值

图 3　掩膜分割过程特征激活值示意图

结合图 2 和图 3 分析可知,掩膜分割操作将原特征图分割为主体和背景两部分,且在两部分中分别增强重点区域的特征表达,抑制另一方的特征表达,分别提取被忽略的背景部分的信息和可能被遗漏的主体部分信息. 掩膜分割操作中的两个掩膜权重矩阵由原特征激活值经通道压缩后正则化得到,额外引入的参数量可忽略不计.

2.4　知识提取模块

2.4.1　注意力分配策略

在训练初期: 分类器网络和自教师网络的分类精度低,难以产生有效的空间注意力,此时过度关注背景知识反而会模糊分类器网络的学习重点,降低学习效率. 而在训练过程中: 对于简单可分样本,分类器网络关注主体即可做出正确判断,对于困难不可分样本,分类器无法依据原有的最后一层特征图得到正确的分类结果,即原有的主体信息无法帮助模型进行准确预测. 因此,从整体和局部两个角度提出动态的注意力分配策略.

从整体而言,在训练初期应降低对背景知识的关注度,随着训练轮次的增加逐步提高对背景知识的关注,在学生网络已具有一定分类精度的基础上帮助学生网络学习更全面的知识. 因此,引入超参数控制对背景知识的整体关注度,其取值范围设置为 $\lambda \in \{0.1, 0.5, 1\}$.

从局部而言,不同样本携带的信息量不同,需要针对每个样本具体分析. 针对简单可分样本,无需提取背景知识便能够预测正确,可减少对背景的关注度; 针对困难不可分样本,凭借原有的单一特征图输入并不能得到正确预测,应提取被忽略的背景知识来辅助分类器网络学习. 此外,存在部分主体区域被错误划分为背景的情况,提取背景知识时也能捕获被遗漏的主体知识.

由上文分析可知,对于简单可分样本,模型使用已有知识即可进行有效预测,可减少对背景知识的学习; 对于困难不可分样本,模型依据已有主体知识无法产生准确的注意力,导致错误预测,应引入背景知识来辅助网络学习,帮助其产生与自教师网络更相似的输出,提高分类精度. 因此,以真实标签为判断基础,基于样本的预测概率分布设计了超参数 τ_b 动态调整背景损失和主体损失的占比,使得正确类别置信度越低、错误置信度越高的样本的背景知识占比越高、主体知识占比越小. 给定真实标签为 y 的输入样本 x, τ_b 的表达式如下所示:

$$\tau_b = \begin{cases} 1, & \text{预测正确;} \\ 2 - \dfrac{\phi(f_y(x;\theta_s)/T)}{\max(\phi(f(x;\theta_s)/T))}, & \text{预测错误.} \end{cases} \tag{4}$$

其中: $f_y(x;\theta_s)$ 表示分类器网络产生的预测类别为 y 的 logit, $f(x,\theta_s)$ 为分类器网络产生的 logit, $\phi(\cdot)$ 为 softmax 函数.

2.4.2　知识提取

知识提取模块以掩膜后的特征图为输入,掩膜分割后的两部分特征图分别经过对应网络的全连接层,

计算得到预测的logit. 令M_b^t和M_b^s分别为自教师和分类器掩膜后的背景特征图,得到自教师网络和分类器网络的背景logit输出$g_t(M_b^t)$和$g_s(M_b^s)$分别为

$$g_t(M_b^t) = W_t M_b^t + b_t,$$

$$g_s(M_b^s) = W_s M_b^s + b_s s. \tag{5}$$

其中:W_t和b_t分别为自教师网络全连接层的权重和偏置,W_s和b_s分别为分类器网络全连接层的权重和偏置.

利用KL损失将忽略的背景知识和丢失的主体知识从自教师网络转移给分类器网络,结合注意力分配策略得到最终的MA损失. 令分类器网络和自教师网络的全连接层分别表示为F_s和F_t,提取背景损失L_b和主体损失L_m的过程如下所示:

$$L_b = \mathrm{KL}\Big(\phi\Big(\frac{g_s(M_b^s)}{T}\Big)\Big\|\phi\Big(\frac{g_t(M_b^t)}{T}\Big)\Big), \tag{6a}$$

$$L_m = \mathrm{KL}\Big(\phi\Big(\frac{g_s(M_m^s)}{T}\Big)\Big\|\phi\Big(\frac{g_t(M_m^t)}{T}\Big)\Big). \tag{6b}$$

其中:M_m^s和M_m^t分别为分类器网络和自教师网络的主体特征图,自教师网络和分类器网络的主体logit输出$g_t(M_m^t)$和$g_s(M_m^s)$的计算方式同式(5),KL为KL散度损失.

2.5 损失函数

如第2.4.1节所述,为了从整体角度控制模型对背景知识的关注度,引入超参数λ. 为了从局部角度调整对背景知识的关注度,设计动态参数τ_b,根据样本携带信息量的不同动态调整主体和背景两部分损失的比例. 结合式(4),得到混合注意力损失L_{MA}的表达式为

$$L_{\mathrm{MA}} = \lambda \cdot (\tau_b \cdot L_b + (2 - \tau_b) \cdot L_m), \tag{7}$$

其中$\lambda \in \{0.1, 0.5, 1\}$且随着训练进程在epoch为100和150时递增.

结合式(4)和(7)分析可知:当样本预测正确时,τ_b恒为1,令分类器网络学习等比例的背景知识和主体知识;而对于预测错误的样本,正确类别的置信度$\phi(f_y(x;\theta_s))$越低,错误预测的置信度$\max(\phi(f(x;\theta_s)/T))$越高,$\tau_b$的值越大,背景损失占总损失的比例越高. 即对于单个样本预测正确时,提供等比例的背景知识和主体知识供分类器学习;预测错误时,对于错误置信度越高的样本,提供的背景知识占比越大,认为原有的主体知识无法指导分类器作出正确预测,需要更多的背景知识来辅助分类器修正错误.

结合MA损失和FRSKD损失,最终总损失为

$$L_{\mathrm{total}} = L_{\mathrm{FRSKD}} + L_{\mathrm{MA}}(x, y, \theta_s, T). \tag{8}$$

3 实 验

3.1 实验设置

3.1.1 数 据 集

本文在4个数据集上对MA-SKD方法进行性能评估. 其中:CIFAR 100[29]和Tiny-ImageNet[30]用于图像分类,Caltech-UCSD Bird (CUB 200)[31]和MIT 67[32]用于细粒度识别任务(fine-grained visual recognition, FGVR). 与传统的图像分类任务相比,FGVR任务具有视觉上更相似的类,每类的训练样本更少.

3.1.2 实施细节

对于CIFAR 100和Tiny-ImageNet数据集,使用ResNet 18[33]和WRN-16-2[34]网络进行实验. 为了使得ResNet 18适应小型数据集,其第1个卷积层被修改为3×3大小、步幅和填充值被修改为1,最大池操作也被移除,Tiny-ImageNet图像的大小已调整为与CIFAR 100相同的大小(32×32). 对于FGVR任务,使用标准的ResNet 18网络.

使用初始学习率为0.1且权重衰减为0.000 1的随机梯度下降算法(SGD)用于优化. 总epoch设置为200,学习率在epoch为100和150时分别减少到之前值的1/10. CIFAR 100和Tiny-ImageNet的批量大小设置为32,FGVR任务的批量大小设置为8. 所有实验均使用标准数据扩充方法,即随机裁剪和翻转.

针对超参数的设置,对于附加数据增强SLA-SD (MA-SKD + SLA)的实验,$\lambda \in \{0.1, 0.5, 1\}$且在训练过程中与学习率进行相同频率的递增. 对于不使用数据增强方法的MA-SKD实验,λ恒为1.

3.2 实验结果

本文对两种实验设置进行实验:MA-SKD为未使用数据增强的方法,MA-SKD + SLA为将所提出方法结合数据增强方法SLA-SD[35]的方法.

在CIFAR 100和Tiny-ImageNet数据集上使用两种分类器网络结构的准确率如表1所示:最优模型加黑表示,次优模型用下划线表示;W 16和R 18分别表示分类器网络为WRN-16-2和ResNet 18;Origin方法表示只使用分类器网络进行实验.

由表1可知:MA-SKD具有比FRSKD更好的性能,使用ResNet 18和WRN-16-2网络在CIFAR 100数据集上的精度分别提升了2.15%和1.54%,使用ResNet 18和WRN-16-2网络在Tiny-ImageNet数据集上的精度分别提升了0.70%和0.57%. 此外,MA-SKD + SLA在使用数据增强的情况下,在CIFAR 100数据集上训练ResNet 18网络的准确率相较于具有

表 1　CIFAR 100与Tiny-ImageNet数据集上的性能比较

方法	年份	CIFAR 100		Tiny-ImageNet	
		W 16	R 18	W 16	R 18
Origin	–	72.07	76.44	51.17	57.93
ONE [36]	2018	73.24	77.33	52.30	57.92
DDGSD [16]	2019	72.01	77.08	51.31	56.7
BYOT [19]	2019	70.48	76.75	50.36	56.91
SAD [37]	2018	70.76	74.98	51.56	54.51
CS-KD [17]	2020	72.47	77.24	50.26	56.56
SLA-SD [35]	2020	73.45	77.82	51.10	58.92
Tf-KD [22]	2021	72.66	77.29	–	56.57
Zipf's LS [23]	2022	–	77.49	–	59.45
AI-KD [24]	2022	–	80.13	–	58.16
MixSKD [18]	2022	75.16	80.45	–	–
DRG+DSR [25]	2023	–	79.30	–	58.08
USKD [26]	2023	–	79.90	–	–
PR-SKD [21]	2024	74.36	78.42	53.32	60.88
FRSKD [20]	2021	73.67	77.51	52.63	59.86
FRSKD + SLA [20]	2021	76.00	81.87	52.20	63.62
MA-SKD	–	75.21	79.66	**53.33**	60.43
MA-SKD+SLA	–	**76.77**	**82.80**	52.51	**63.78**

相同数据增强设置的FRSKD + SLA提高了0.93 %, 在Tiny-ImageNet数据集上训练WRN-16-2网络的准确率提高了0.31 %. 综上所述, MA-SKD方法在图像分类数据集(CIFAR 100和Tiny-ImageNet)上平均精度提升了1.24 %, MA-SKD和MA-SKD + SLA方法取得了4个实验上的最优精度.

分析表1可知, MA-SKD方法在CIFAR 100数据集上的精度提升高于Tiny-ImageNet(下文简称TINY)数据集上的精度提升. 对比MA-SKD与Origin可知: MA-SKD在CIFAR 100和TINY数据集上的平均精度相比于只采用WRN-16-2和ResNet 18作为分类器网络的Origin方法分别提升了3.18 %和2.33 %, 表明自教师网络能够提供有效的精细化特征图且MA-SKD方法能够进一步帮助分类器网络逼近自教师网络输出, 提升分类性能.

表2为所提出方法在CUB 200和MIT 67数据集上训练ResNet 18网络的性能. 最优模型加黑表示, 次优模型用下划线表示. Origin方法表示只使用分类器网络进行实验.

由表2可知: MA-SKD方法相较于基线模型FRSKD在CUB 200和MIT 67数据集上的准确率分别提升了4.78 %和0.55 %, 在FGVR任务(CUB 200和MIT 67)上平均精度提升了2.67 %, 而使用数据增强的MA-SKD + SLA比使用相同数据增强设置的

表 2　CUB 200与MIT 67数据集上的性能比较

方法	年份	CUB 200	MIT 67
Origin	–	52.43	61.41
ONE [36]	2018	55.13	57.52
DDGSD [16]	2019	59.04	59.77
BYOT [19]	2019	59.17	59.12
SAD [37]	2018	53.33	55.78
CS-KD [17]	2020	64.42	57.73
SLA-SD [35]	2020	56.88	62.63
Tf-KD [22]	2021	64.08	61.92
AI-KD [24]	2022	70.41	63.16
MixSKD [18]	2022	**72.15**	64.10
PR-SKD [21]	2024	64.95	66.82
FRSKD [20]	2021	64.82	63.73
FRSKD + SLA [20]	2021	66.56	67.08
MA-SKD	–	69.60	64.28
MA-SKD + SLA	–	70.07	**68.13**

FRSKD + SLA在CUB 200和MIT 67数据集上的准确率分别提升了3.51 %和1.05 %. MA-SKD方法在CUB 200和MIT 67数据集上相比于只采用ResNet 18作为分类器网络的Origin方法平均精度分别提升了17.17 %和2.87 %, 可知自教师网络提供的融合低层位置信息和高层语义信息的特征图对CUB 200鸟类数据集的提升效果明显优于MIT 67室内数据集. 分析认为仅使用ResNet 18网络可能无法有效提取CUB 200数据集中鸟类的细节判别特征.

3.3　消融实验

3.3.1　超参数

MA-SKD + SLA方法使用不同的λ取值的性能比较如表3所示, 其中方法a～方法e分别对应λ的取值为$\lambda = 0.1$、$\lambda = 0.5$、$\lambda = 1$、$\lambda \in \{0.5, 0.7, 1\}$和$\lambda \in \{0.1, 0.5, 1\}$. 最优模型加黑表示, 次优模型用下划线表示.

表 3　不同λ取值的性能比较

方法	CIFAR 100		TINY		CUB 200	MIT 67
	W 16	R 18	W 16	R 18	R18	R 18
a	76.56	82.78	50.65	63.70	67.82	67.74
b	76.49	80.72	51.36	62.57	67.55	67.38
c	**76.80**	82.68	51.67	63.74	67.55	67.68
d	76.69	82.78	52.06	63.67	67.78	**70.14**
e	76.77	**82.80**	**52.51**	**63.78**	**70.07**	68.13

由表3可知: 在TINY数据集和CUB 200数据集的实验中, 方法e表现最优; 在MIT 67数据集和使用WRN-16-2网络的CIFAR 100数据集上, 方法e具有次

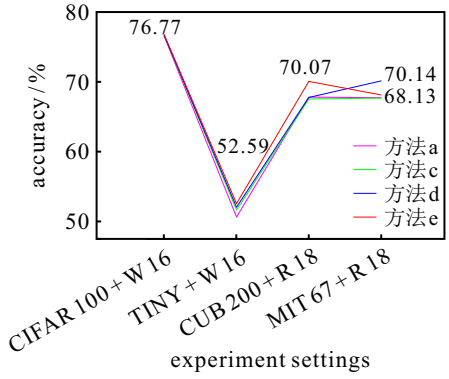
优的性能; 方法 d 虽然在 MIT 67 数据集上表现最佳, 但是在其他实验中的表现一般. 由此可知, MASKD + SLA 的参数设置采用方法 e 可得到最优表现, 即 $\lambda \in \{0.1, 0.5, 1\}$.

针对 τ_b 的不同初始值进行消融实验, 令 $x \in \{1, 2, 3\}$, 其表达式如下所示:

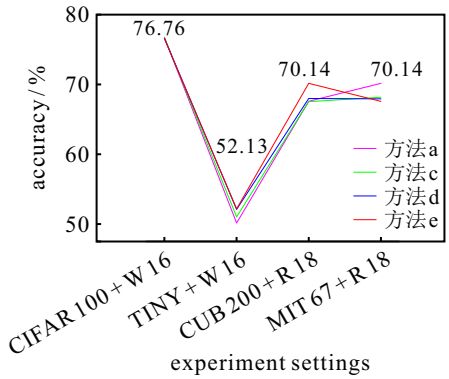
$$\tau_b = \begin{cases} x, & \text{预测正确;} \\ x + 1 - \dfrac{\phi(f_y(x;\theta_s)/T)}{\max(\phi(f(x;\theta_s)/T))}, & \text{预测错误.} \end{cases} \tag{9}$$

式 (7) 中对应 L_M 的权重为 $x + 1 - \tau_b$

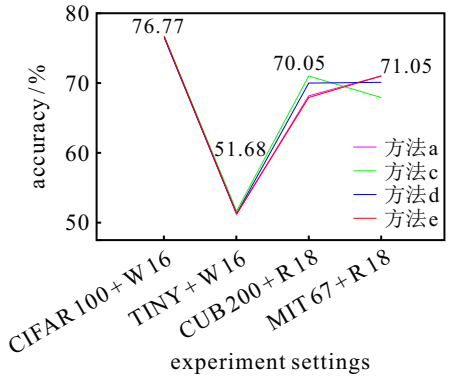
针对 λ 和 x 的消融实验结果如图 4 所示. 其中: 图 4(a) ~ 图 4(c) 对应 x 的不同取值; 方法 a、方法 c ~ 方法 e 与表 3 的实验设置相同, 分别对应 λ 的取值为 $\lambda = 0.1$、$\lambda = 1$、$\lambda \in \{0.5, 0.7, 1\}$ 和 $\lambda \in \{0.1, 0.5, 1\}$. 由图 4 可知, $x = 1$ 和 $\lambda \in \{0.1, 0.5, 1\}$ 为最佳实验设置, 在 4 个实验中具有稳定的优越表现.



(a) $x = 1$



(b) $x = 2$



(c) $x = 3$

图 4 4 个数据集上的消融实验结果

3.3.2 混合注意力

使用背景注意力方法与不同比例混合注意力方法的性能对比如表 4 所示: W 16 和 R 18 分别表示分类器网络为 WRN-16-2 和 ResNet 18; 方法 A 为不关注背景的 FRSKD + SLA; 方法 B 为额外引入背景损失的 MA-SKD + SLA; 方法 C 为额外引入 1 : 1 固定比例的主体和背景损失的 MA-SKD + SLA; 方法 D 为使用动态的主体和背景注意力分配策略的 MA-SKD + SLA, 4 者具有相同的初始参数设置. 最优模型加黑表示, 次优模型用下划线表示.

表 4 不同注意力方法的性能比较

方法	CIFAR 100		TINY		CUB 200	MIT 67
	W 16	R 18	W 16	R 18	R 18	R 18
A	76.00	81.87	<u>52.20</u>	<u>63.62</u>	66.56	67.08
B	76.44	81.92	51.96	62.05	68.34	**68.20**
C	<u>76.53</u>	<u>82.61</u>	51.29	62.66	<u>69.46</u>	67.98
D	**76.77**	**82.80**	**52.51**	**63.78**	**70.07**	<u>68.13</u>

分析表 4 可知: 在相同实验条件的 4 个数据集上, 使用动态混合注意力的方法 D 表现优越, 在多个数据集上均表现出最高精度; 而引入背景注意力的方法 B 和引入混合注意力的方法 C 在多数实验上均优于不关注背景的方法 A, 且方法 C 的整体表现优于仅使用背景注意力的方法 B. 由此可知: 提取背景知识有助于模型学习, 且引入混合注意力能够均衡背景知识; 设计动态的注意力分配策略能够针对携带信息量不同的样本调整混合注意力的比例, 提升分类器网络性能.

3.4 可 视 化

CIFAR 100 和 CUB 200 上部分图像的注意力图可视化如图 5 所示. 由图 5 可知: MA-SKD 方法挖掘前背景知识并引导其相互协作, 促使分类器网络产生更精准的注意力, 因此, 相较于基准网络 FRSKD 具有更好的注意力可视化结果, 具体分析如下.

由图 5 第 1 列和第 2 列可知: FRSKD 错误地关注了大范围的橘色背景和蓝色海面; 而 MA-SKD 对瓢虫和鲸鱼均具有很强的识别能力, 能够准确聚焦在瓢虫和鲸鱼上, 不受相似颜色背景的干扰. 由图 5 第 3 列和第 4 列可知: FRSKD 仅关注鸟的中间部分, 忽略了尾巴、翅膀等细节特征; 而 MA-SKD 具有连续完整的注意力图, 能够准确识别出翅膀和杂乱背景中的黑色尾巴, 表明 MA-SKD 对主体部分有更强的识别能

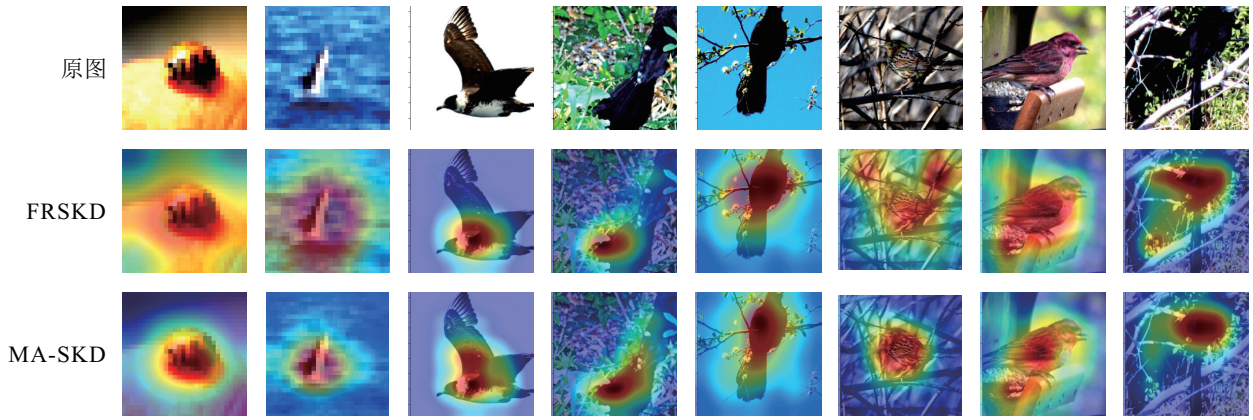


图 5　CIFAR 100 与 CUB 200 上的注意力对比

力. 分析图 5 第 5 列可知: FRSKD 产生的注意力图存在中心偏移的情况, 而 MA-SKD 能够改善注意力偏移并准确聚焦. 由图 5 第 6 列～第 8 列可知: FRSKD 的注意力被相似颜色的背景所分散, 而 MA-SKD 能够减少对相似背景的错误关注, 更精准地聚焦主体.

4　结　论

本文提出了基于动态混合注意力的自知识蒸馏方法 MA-SKD. 所提出方法通过提取遗漏的主体信息和被忽略的背景知识辅助分类器网络学习. 首先, 本文设计了掩膜分割模块将特征图分割为主体特征图和背景特征图, 分别学习主体知识和背景知识作为补充, 改善了原注意力划分区域过大或过小的问题; 然后, 本文提出了基于动态注意力分配策略的知识提取模块, 在训练过程中逐步增加对背景知识的关注, 且针对携带信息量不同的样本动态调整主体损失和背景损失的占比. MA-SKD 方法能够捕获遗漏的主体知识和被忽略的背景知识供分类器网络学习, 在图像分类数据集 CIFAR 100 和 TINY 上平均精度较 FRSKD 提升了 1.24 %, 在 FGVR 数据集 CUB 200 和 MIT 67 上平均精度较 FRSKD 提升了 2.67 %, 表明所提出方法在图像分类和 FGVR 任务上均表现优异. 然而, 所提出方法的掩膜分割模块依赖于自教师网络提供的精确注意力图, 如何在在线蒸馏等场景下提取背景知识值得进一步研究.

参考文献(References)

[1] Gou J P, Yu B S, Maybank S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[2] Sultana F, Sufian A, Dutta P. Advancements in image classification using convolutional neural network[C]. The 4th International Conference on Research in Computational Intelligence and Communication Networks. Kolkata, 2018: 122-129.

[3] Zou Z X, Chen K Y, Shi Z W, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023, 111(3): 257-276.

[4] 金超熊. 基于知识蒸馏的输电线路螺栓缺陷图像分类[D]. 北京: 华北电力大学, 2021.
(Jin C X. Bolt defect image classification of transmission line based on knowledge distillation[D]. Beijing: North China Electric Power University, 2021.)

[5] 王磊, 乔俊飞, 杨翠丽, 等. 基于灵敏度分析的模块化回声状态网络修剪算法[J]. 自动化学报, 2019, 45(6): 1136-1145.
(Wang L, Qiao J F, Yang C L, et al. Pruning algorithm for modular echo state network based on sensitivity analysis[J]. Acta Automatica Sinica, 2019, 45(6): 1136-1145.)

[6] Idelbayev Y, Carreira-Perpiñán M Á. Optimal selection of matrix shape and decomposition scheme for neural network compression[C]. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, 2021: 3250-3254.

[7] Qin H T, Gong R H, Liu X L, et al. Binary neural networks: A survey[J]. Pattern Recognition, 2020, 105: 107281.

[8] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: Inverted residuals and linear bottlenecks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 4510-4520.

[9] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, 14(7): 38-39.

[10] 潘瑞东, 孔维健, 齐洁. 基于预训练模型与知识蒸馏的法律判决预测算法[J]. 控制与决策, 2022, 37(1): 67-76.
(Pan R D, Kong W J, Qi J. Legal judgment prediction based on pre-training model and knowledge distillation[J]. Control and Decision, 2022, 37(1): 67-76.)

[11] 程旗, 李捷, 高晓利, 等. 基于深度稀疏低秩分解的深度神经网络轻量化方法[J]. 控制与决策, 2023, 38(3): 751-758.
(Cheng Q, Li J, Gao X L, et al. Lightweight method of deep neural network based on deep sparse low rank

decomposition[J]. Control and Decision, 2023, 38(3): 751-758.)

[12] Su C P, Tseng C H, Lee S J. Knowledge from the dark side: Entropy-reweighted knowledge distillation for balanced knowledge transfer[J/OL]. 2023, arXiv: 2311.13621.

[13] Li X W, Lin L, Wang S, et al. Unlock the power: Competitive distillation for multi-modal large language models[J/OL]. 2023, arXiv: 2311.08213.

[14] Li S J, Lin M B, Wang Y, et al. Distilling a powerful student model via online knowledge distillation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(11): 8743-8752.

[15] Choi J, Cho H, Cheung S, et al. ORC: Network group-based knowledge distillation using online role change[C]. IEEE/CVF International Conference on Computer Vision. Paris, 2023: 17381-17390.

[16] Xu T B, Liu C L. Data-distortion guided self-distillation for deep neural networks[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 5565-5572.

[17] Yun S, Park J, Lee K, et al. Regularizing class-wise predictions via self-knowledge distillation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 13876-13885.

[18] Yang C G, An Z L, Zhou H L, et al. MixSKD: Self-knowledge distillation from mixup for image recognition[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 534-551.

[19] Zhang L F, Song J B, Gao A N, et al. Be your own teacher: Improve the performance of convolutional neural networks via self distillation[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 3713-3722.

[20] Ji M, Shin S, Hwang S, et al. Refine myself by teaching myself: Feature refinement via self-knowledge distillation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 10664-10673.

[21] Yu H, Feng X, Wang Y L. Enhancing deep feature representation in self-knowledge distillation via pyramid feature refinement[J]. Pattern Recognition Letters, 2024, 178: 35-42.

[22] Yuan L, Tay F E, Li G L, et al. Revisiting knowledge distillation via label smoothing regularization[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 3903-3911.

[23] Liang J, Li L, Bing Z, et al. Efficient one pass self-distillation with zipf's label smoothing[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 104-119.

[24] Kim H, Suh S, Baek S, et al. AI-KD: Adversarial learning and Implicit regularization for self-knowledge distillation[J/OL]. 2022, arXiv: 2211.10938.

[25] Wang X C, Han P C, Guo L. Lightweight self-knowledge distillation with multi-source information fusion[J/OL]. 2023, arXiv: 2305.09183.

[26] Yang Z D, Zeng A L, Li Z, et al. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels[J/OL]. 2023, arXiv: 2303.13005.

[27] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 10778-10787.

[28] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8759-8768.

[29] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[J]. Handbook of Systemic Autoimmune Diseases, 2009, 1(4): 1-60.

[30] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]. IEEE Conference on Computer Vision and Pattern Recognition. Miami, 2009: 248-255.

[31] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[J]. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[32] Quattoni A, Torralba A. Recognizing indoor scenes[C]. IEEE Conference on Computer Vision and Pattern Recognition. Miami, 2009: 413-420.

[33] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 770-778.

[34] Zagoruyko S, Komodakis N. Wide residual networks[J/OL]. 2016, arXiv: 1605.07146.

[35] Lee H, Hwang S J, Shin J. Self-supervised label augmentation via input transformations[C]. Proceedings of the 37th International Conference on Machine Learning. New York, 2020: 5714-5724.

[36] Lan X, Zhu X T, Gong S G. Knowledge distillation by on-the-fly native ensemble[J/OL]. 2018, arXiv: 1806.04606.

[37] Miyato T, Maeda S I, Koyama M, et al. Virtual adversarial training: A regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979-1993.

作者简介

唐媛 (2000−), 女, 硕士生, 主要研究方向为知识蒸馏, E-mail: 835439442@qq.com;

陈莹 (1976−), 女, 教授, 博士生导师, 主要研究方向为计算机视觉、模式识别、多媒体信息融合、深度模型压缩, E-mail: chenying@jiangnan.edu.cn.