基于通道相关性的类注意力知识蒸馏

吴华涛,朱子奇
(武汉科技大学 计算机科学与技术学院,湖北 武汉 430065)

摘 要:以前的知识蒸馏方法在模型压缩上表现出了令人印象深刻的性能,其中在基于类注意力转移的知识蒸馏(CAT-KD)这项工作中证明了通过转移类激活图可以使学生模型获得和增强识别输入分类区域的能力,这种能力是当前主流 CNN 模型进行分类的关键。其通过将类激活图平均池化和 l2 归一化的方式来转移类激活图,从而增强学生模型识别输入分类区域的能力,提高蒸馏性能。然而,这种方式忽略了类激活图中的通道相关的知识,这对于学生模型学习识别输入分类区域的能力至关重要。为了解决上述问题,该文提出了基于通道相关性的类注意力转移方法。具体来说,为了从类激活图中提取丰富的知识,该方法不仅考虑了样本内的类激活图中不同通道的特征知识,还考虑了不同样本的类激活图中基于每通道特征的关系知识。实验表明,该方法在 CIFAR-100 数据集上比基准方法提升了 0.96 百分点,优于对比方法。

关键词:知识蒸馏;通道相关性;类激活图;通道知识;注意力蒸馏
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1673-629X(2024)12-0125-07
doi:10.20165/j.cnki.ISSN1673-629X.2024.0242

Class Attention Knowledge Distillation Based on Channel Correlation

WU Hua-tao,ZHU Zi-qi
(School of Computer Science & Technology,Wuhan University of Science and Technology,Wuhan 430065,China)

Abstract:Previous knowledge distillation methods have shown impressive performance in model compression,among which in the work of Class Attention Transfer Based Knowledge Distillation (CAT-KD),it has been proven that the transfer class activation graph can enable student models to acquire and enhance the ability to recognize input classification regions,which is the key to current mainstream CNN models for classification. It enhances the ability of the student model to recognize input classification regions and improves distillation performance by transferring class activation maps through average pooling and l2 normalization. However,this approach ignores the channel related knowledge in the class activation maps,which is crucial for the student model's ability to learn and recognize input classification regions. To address the aforementioned issues,we propose a class attention transfer method based on channel correlation. Specifically,in order to extract rich knowledge from class activation maps,the proposed method not only considers the feature knowledge of different channels in the class activation maps within the samples,but also considers the relationship knowledge based on each channel feature in the class activation maps of different samples. The experiment shows that the proposed method has improved by 0.96 percentage points compared to the benchmark method on the CIFAR-100 dataset,which is better than the comparison method.

Key words:knowledge distillation;channel correlation;class activation maps;channel knowledge;attention distillation

0 引 言

随着深度神经网络在计算机视觉识别方面取得的巨大成功,在许多视觉识别任务中使用层次更深、参数更多的深度神经网络来提取丰富的语义信息。但随之而来的是需要大量的计算和内存资源,这使得在资源有限的移动设备或嵌入式系统上部署深度学习模型十分困难。为了解决这一问题,模型压缩技术越来越受到深度学习界的关注,主要包括量化、剪枝和知识蒸馏[1]。其中,知识蒸馏的主要思想是从较大的教师网络中提取知识来提升学生网络的性能。根据转移知识的类型,知识蒸馏可以分为三类:基于响应[2]、基于特征[3-7]、基于注意力[8-12]。

在基于注意力的方法中,前人的研究 AT[8]验证了注意转移的有效性,首次提出了通过转移注意力来进行知识蒸馏。Guo 等人提出的 CAT-KD[9]通过转移类激活图(Class Activation Maps,CAM)[13]来使学生

收稿日期:2024-03-05 **修回日期**:2024-07-09
基金项目:公安部科技计划项目(2022JSM08)
作者简介:吴华涛(2000-),男,硕士研究生,研究方向为计算机视觉;通信作者:朱子奇(1983-),男,博士,副教授,CCF 会员(55349M),研究方向为计算机视觉、模式识别。

模型学习到教师模型识别输入的类别区分区域的能力。如图 1 所示,将主流模型的结构稍加转换,就可以得到 CAM,它是一种表明特定类别输入的判别区域的类注意图。可以看出,CAMs 中的每一个通道都代表着一个类激活图,每一个通道中都包含着大量的特征知识。然而,在 CAT-KD 中并没有考虑到 CAMs 通道内知识以及不同实例间 CAMs 通道间的关系知识。

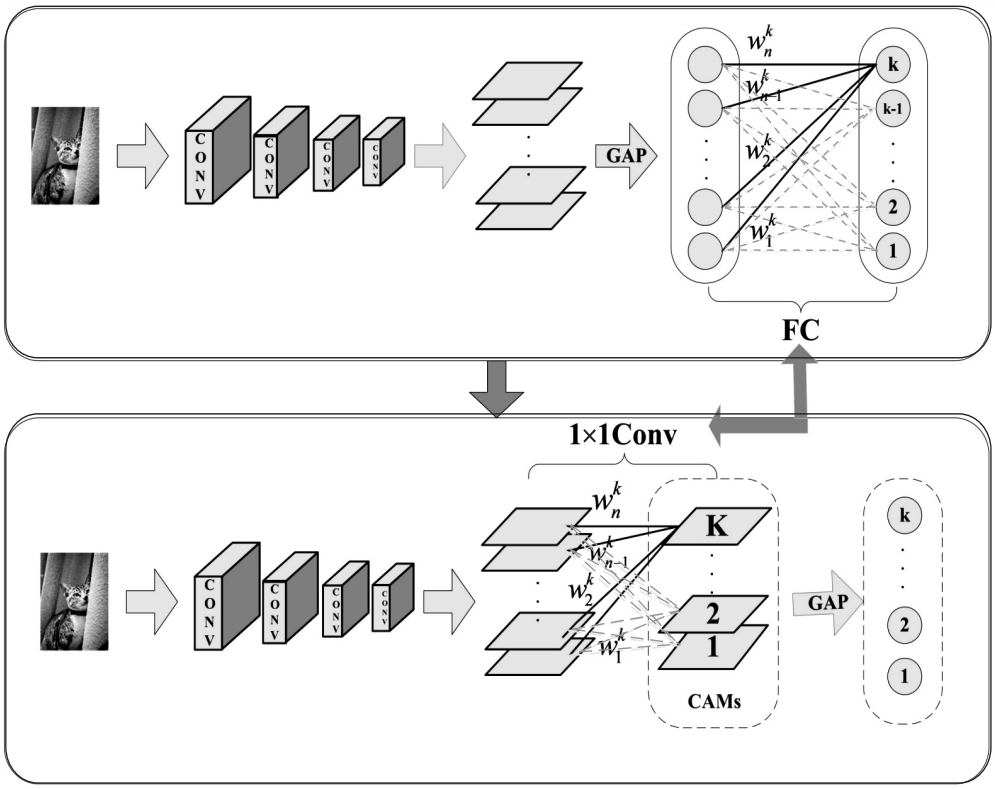


图 1 将传统的 CNN 转换后的结构示意图

为了解决上述问题,该文提出了基于通道相关性的类注意力知识蒸馏。具体来说,为了从 CAMs 中获取丰富的知识,考虑了基于个体实例和实例关系的 CAMs 通道知识。与现有的知识蒸馏方法相比,该方法可以充分转移 CAMs 的通道特征知识,从而增强学生模型识别输入的类别区分区域的能力。

综上所述,主要贡献如下:

(1)考虑到实例间关系,设计了一种基于 CAMs 通道特征的关系知识转移方案。

(2)通过将转移实例内 CAMs 通道特征知识和转移实例间 CAMs 通道特征的关系知识相结合,提出了一种新的基于注意力的知识蒸馏方法,即基于通道相关性的类注意力转移。

(3)通过充分的实验表明,该方法在图像分类数据集(CIFAR-100、ImageNet)上有效地提高了蒸馏性能,与多个知识蒸馏方法相比,该方法取得了最优性能。

1 相关工作

知识蒸馏这一概念最早是由 Hinton[14] 于 2015 年提出。作为一种迁移学习方法,知识蒸馏旨在通过将较大的教师网络中提取的暗知识迁移到较小的学生网络中来提高学生网络的性能。以前的知识蒸馏根据转移知识的类型可以分为三类:基于响应、基于特征和基于注意力。

在先前的工作中,AT 是第一种基于转移注意的知识蒸馏方法,它将注意图定义为表明模型最关注的输入区域的空间图。在实践中,通过计算特征图的和来得到注意力图,然后转移得到的注意力图来提高学生网络的性能。而在 CAT-KD 这一工作中,作者提出了通过转移类激活图(Class Activation Maps,CAM)来提高学生网络的性能。类激活图这一概念由 Zhou 等人[10] 提出,利用高级特征图和全连接层的参数来生成特定类别的注意图,如图 1 所示。将 FC 层转换成具有 1×1 核的卷积层,其中卷积层的权重为 FC 层的权重,并移动全局平均池化层的位置,在前向传播过程中得到 CAMs。可以看出,CAMs 的每一个通道都代表着某一个类别的注意力图。在上述工作中,转移 CAMs 的过程是通过 MSE 损失精确匹配教师模型和学生模型的 CAMs,它忽略了 CAMs 的通道知识以及实例间关系的知识。在本节中将简要讨论使用通道和关系知识进行转移 CAMs 的相关工作。

1.1　通道知识

近年来,来自教师网络中间层的通道特征知识被作为一种提炼出来的知识用于学习学生网络。例如,Zhao 等人[15]提出了通道蒸馏(CD),将通道信息提炼为从教师网络到学生网络的通道关注。Qu 等人[16]提出了混合注意转移,该方法从教师网络中的通道和空间语义信息中提取知识。Shu 等人[17]提出了一种基于通道的蒸馏方法,该方法使学生网络模拟由通道特征图计算的教师软分布。Yoon 等人[18]提出了相似性保持知识蒸馏(SPKD),它将通道、空间和批量单位上的成对相似性从教师提炼到学生。

通道特征知识通常是针对单个实例来考虑的,而从通道特征的角度考虑来自实例关系的知识则很少。在该文提出的方法中通过结合实例内的通道知识和不同实例间的通道关系知识来从教师网络中提炼 CAMs 到学生网络中。

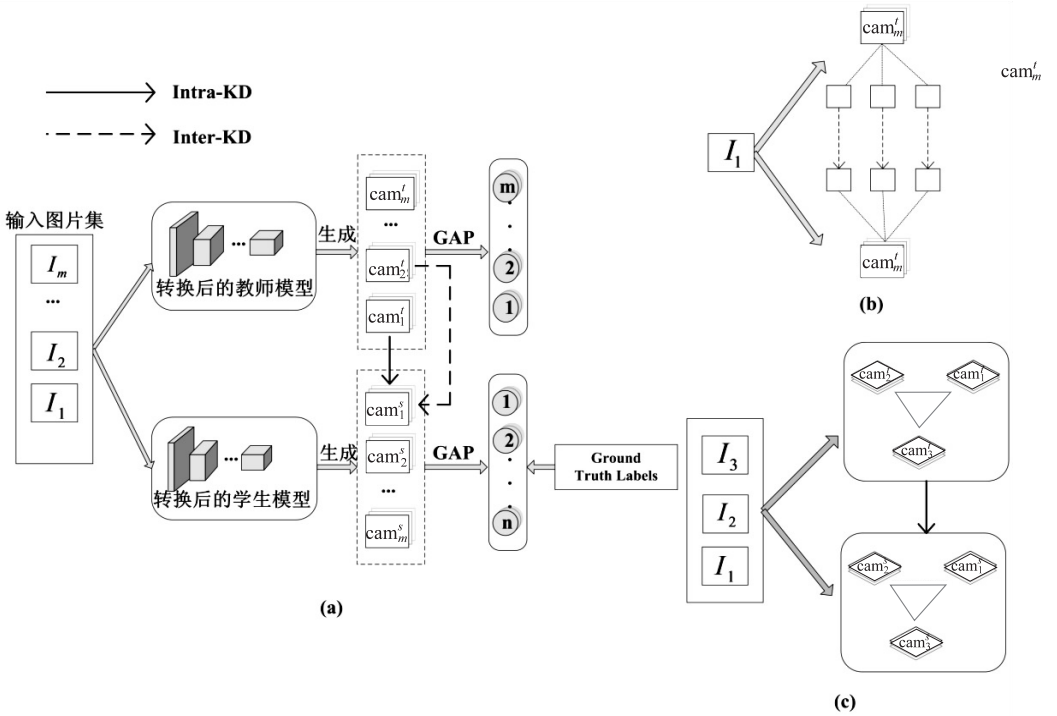
1.2　基于关系的知识

近年来,越来越多的知识蒸馏方法探索实例间存在的关系知识。例如,Park 等人[5]提出了一种关系知识蒸馏(RKD),它简单地使用角度和距离方向的蒸馏损失来衡量实例关系的知识。Liu 等人[19]提出了一种基于实例关系图的知识蒸馏,该方法从单个实例、实例关系和特征空间变换中对特征知识进行建模。Passalis 和 Tefas[20]从数据概率分布的角度提出了概率知识转移。Peng 等人[21]同时考虑了实例同余和实例间同余,注意到了实例间的相关性。Ye 等人[22]提出的关系促进了局部分类器蒸馏,使学生匹配来自老师的实例之间的相似性。Ge 等人[23]在师生架构上开发了混合阶 RKD,用于提高低分辨率图像识别。Zhu 等人[24]提出了互补关系对比蒸馏,进一步转移实例间丰富的知识。Huang 等人[25]提出了从一个更强的教师中进行蒸馏,以保留教师和学生之间预测的类间和类内关系知识。Guo 等人提出了从实例内和实例间的通道相关中提取知识。

基于上述工作的启发,实例间的关系信息作为教师网络的结构知识,对学习高效的学生网络非常有用。因此,该文提出了一种新的类注意力转移方法,通过提炼 CAMs 实例内的通道知识和实例间关系知识来更为高效地转移 CAMs,从而增强学生模型识别输入的类别区分区域的能力,提高蒸馏性能。

2　文中方法

在本节中,详细介绍了基于通道相关性的类注意力知识蒸馏方法,总体框架如图 2 所示。该方法主要包括两个部分:(1)实例内(Intra)CAMs 通道蒸馏,如图 2(b)所示;(2)实例间(Inter)CAMs 通道相关性蒸馏,如图 2(c)所示。



(a)模型架构;(b)实例内 CAMs 通道蒸馏;(c)实例间 CAMs 通道关系蒸馏

图 2　文中方法概述

2.1　回顾 CAMs 的生成

在图像分类任务中,主流模型通常通过 CNN 提取特征,然后将提取到的高级特征通过全局平均池化送到全连接层得到 logit。设 $F=[F_1,F_2,\cdots,F_C]\in$

$\mathbb{R}^{C\times W\times H}$ 表示最后一个卷积层生成的特征图,其中 C、W、H 分别代表通道数量、宽度、高度。$f_j(x,y)$ 表示空间位置 (x,y) 在 j 通道中 F 的激活,GAP 为全局平均池化层。

则主流模型 logits 的计算过程可以写成:

$$L_i = \sum_{1\leqslant j\leqslant C} w_j^i \times \mathrm{GAP}(F_j) = \frac{1}{W\times H}\sum_{x,y}\sum_{1\leqslant j\leqslant C} w_j^i \times f_j(x,y) \tag{1}$$

其中,L_i 表示第 i 类的 logit,w_j^i 为经过全局平均池化后第 i 类对应的全连接层(FC 层)的权重。根据先前的工作[13],可以得到第 i 类对应 CAM 的计算过程:

$$\mathrm{CAM}_i(x,y) = \sum_{1\leqslant j\leqslant C} w_j^i \times f_j(x,y) \tag{2}$$

如图 1 所示,将 FC 层转换为 1×1 的卷积层,并移动 GAP 层的位置。通过公式 1 和公式 2 可以看出,转换后的模型生成 L_i 的过程可由下式得到:

$$L_i = \frac{1}{W\times H}\sum_{x,y}\left[\sum_{1\leqslant j\leqslant C} w_j^i \times f_j(x,y)\right] = \frac{1}{W\times H}\sum_{x,y}\mathrm{CAM}_i(x,y) = \mathrm{GAP}(\mathrm{CAM}_i) \tag{3}$$

从公式 3 可以看出,转换后的模型的 logit 可以通过计算 CAM 的平均池化得到。在 CAMs 中拥有识别输入的类别区分区域的能力,这对于分类任务至关重要。为了更高效地转移 CAMs,提高蒸馏性能,该文提出了通过蒸馏实例内 CAMs 的通道知识和实例间 CAMs 的通道关系知识在师生网络中提炼 CAMs,增强学生网络识别输入的类别区分区域的能力。

2.2 实例内 CAMs 通道蒸馏

在本节中,为了将信息丰富的 CAMs 实例特征知识从教师网络中提取到学生网络中,该文采用对齐师生网络 CAMs 通道权重的方法来实现师生网络之间的通道一致性。设 c_T^{ij} 和 c_S^{ij} 分别表示教师网络和学生网络中第 i 个实例的 CAMs 第 j 个通道的权重。则权重 c_T^{ij} 通过全局平均池化的计算如下:

$$c_T^{ij} = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathrm{cam}(h,w) \tag{4}$$

其中,H 和 W 表示特征映射的空间维度大小,$\mathrm{cam}(h,w)$ 表示 CAM 在空间位置 (h,w) 的激活。在这里通道权重代表了通道信息统计,它增强了 CAMs 通道信息特征,同时弱化了 CAMs 中不必要的通道特征。因此,通过通道权重的方式从 CAMs 中获得的知识更有价值。通过通道权重的制定,实例内通道蒸馏损失的计算过程如下:

$$L_{\mathrm{Intra}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k}(c_T^{ij}-c_S^{ij})^2}{m\times k} \tag{5}$$

其中,m 表示批量大小,k 表示通道数,L_{Intra} 为实例内的 CAMs 通道蒸馏损失。如公式 5 所示,通过最小化教师网络和学生网络的 CAMs 通道权重之间的差异,可以很好地传递实例内 CAMs 通道特征知识,提高蒸馏性能。

2.3 实例间 CAMs 通道相关性蒸馏

为了进一步提高 KD 性能,从教师网络中挖掘实例间的 CAMs 通道关系知识来学习学生网络。在以往的 RKD 方法中,不同实例之间的关系总是通过深度模型的特征映射来建模。为了获取更丰富的实例间关系知识,该文从通道的角度重新提取了实例间关系知识,将实例间的关系知识通过不同实例间 CAMs 的通道相关性来体现。

提出的实例间 CAMs 通道关系蒸馏主要通过两个方面来体现:(1)基于 CAMs 通道距离方向的蒸馏损失;(2)基于 CAMs 通道角度方向的蒸馏损失。基于以上两个方面,可以得到实例间 CAMs 通道相关蒸馏。设 x^n 为不同实例的第 j 个通道权重的 n 个元组的集合。例如,2 元组和 3 元组对应的集合分别记为 $x^2=\{(c^{uj},c^{vi})\mid u\neq v\}$ 和 $x^3=\{(c^{uj},c^{vj},c^{wj})\mid u\neq v\neq w\}$。$c^{uj}$ 代表网络中第 u 个实例的第 j 个通道权重,其定义与公式 4 相同。则任意两个实例之间的通道相关性定义如下:

$$l(c^{uj},c^{vj}) = \frac{1}{\pi}\parallel c^{uj}-c^{vj}\parallel^2 \tag{6}$$

其中,$l(c^{uj},c^{vj})$ 代表 CAMs 中第 u 个实例和第 v 个实例的第 j 个通道之间的距离关系。π 是一个归一化因子,它的计算公式为:

$$\pi = \frac{1}{|x^2|}\sum_{(c^u,c^v)\in x^2}\parallel c^{uj}-c^{vj}\parallel^2 \tag{7}$$

其中,$|x^2|$ 表示 2 元组的数量,x^2 代表所有 2 元组的集合。在蒸馏过程中,学生网络中的跨实例通道的距离相关性尽可能地与教师网络中的相匹配。通过这种方式,可以得到师生网络实例间 CAMs 通道相关性的距离蒸馏损失公式如下:

$$L_{\mathrm{DCRD}}(c^u,c^v) = \sum_{(c^u,c^v)\in x^2}\sum_{j=1}^{k} H(l(c_T^{uj},c_T^{vj}),l(c_S^{uj},c_T^{vj})) \tag{8}$$

其中,$c^u=\{c^{u1},c^{u2},\cdots,c^{uk}\}$ 代表神经网络中第 u 个实例的 k 个通道权值的集合,c_T^{uj} 和 c_S^{uj} 分别表示教师网络和学生网络的第 u 个实例的第 j 个通道权值,$H(.)$ 代表 Huber 损失,其计算过程如下:

$$H(p,q) = \begin{cases}\dfrac{1}{2}(p-q)^2, & \text{if } |p-q|\leqslant 1\\ |p-q|-\dfrac{1}{2}, & \text{otherwise}\end{cases} \tag{9}$$

任意三个实例的 CAMs 之间的通道相关性由角度方向损失的定义如下:

$$r(c^{uj},c^{vj},c^{wj})=\cos\angle c^{uj}c^{vj}c^{wj}=\langle e^{uvj},e^{wvj}\rangle \tag{10}$$

其中, $e^{uvj}=(c^{uj}-c^{vj})/\|c^{uj}-c^{vj}\|_2$, $e^{wvj}=(c^{wj}-c^{vj})/\|c^{wj}-c^{vj}\|_2$,则师生网络之间的实例间 CAMs 通道相关性的角度蒸馏损失公式如下:

$$L_{\mathrm{ACRD}}(c^u,c^v,c^w)=\sum_{(c^u,c^v,c^w)\in x^3}\sum_{j=1}^{k}H(r(c_T^{uj},c_T^{vj},c_T^{wj}),r(c_S^{uj},c_S^{vj},c_S^{wj})) \tag{11}$$

在该文提出的方法中,同时考虑了距离和角度的蒸馏损失。因此,结合公式 8 中的距离损失和公式 11 中的角度损失,可以得到实例间 CAMs 通道相关性的蒸馏损失定义:

$$L_{\mathrm{Inter}}=L_{\mathrm{DCRD}}(c^u,c^v)+\eta L_{\mathrm{ACRD}}(c^u,c^v,c^w) \tag{12}$$

其中, η 是用来平衡项的参数。可以看出。通过通道加权机制表征实例间 CAMs 通道相关性,加强了不同实例 CAMs 中重要通道特征之间的关系知识,同时削弱了 CAMs 中非必要通道特征之间的关系知识。在蒸馏过程中,通过在师生网络中传递 CAMs 通道特征之间的关系知识,增强了学生模型识别输入的类别区分区域的能力,提高了学生的分类性能。

2.4　损失函数

文中方法同时考虑了实例内 CAMs 的通道一致性和实例间 CAMs 的通道相关性在师生网络中转移 CAMs。由公式 5、公式 12 可以得到该方法的总损失函数定义:

$$L_{\mathrm{KD}}=\alpha L_{\mathrm{Intra}}+\beta L_{\mathrm{Inter}}+\gamma L_{\mathrm{CE}} \tag{13}$$

其中, α、β 和 γ 代表平衡超参数, L_{CE} 为交叉熵损失函数。

3　实验与分析

3.1　数据集

CIFAR-100 是一个包含 100 个类别的图像数据集,每个类别包含 600 张尺寸为 32×32 的彩色图像。数据集中的图像被均匀分为训练集和测试集,每个集合中有 50 000 张图像和 10 000 张图像。这些图像涵盖了各种物体和场景,例如,动物、植物、交通工具、家具等。每个图像都有一个标签,表示它所属的类别。

ImageNet 是一个庞大的图像数据集,包含数百万张高分辨率图像,涵盖了数千个类别。该文使用的版本是 ImageNet Large Scale Visual Recognition Challenge(ILSVRC)数据集,其中包含 1 000 个类别,每个类别大约有 1 000 张图像。

3.2　实验参数

对于 CIFAR-100 数据集,实验的 Batchsize 大小为 64,使用 SGD 对所有模型训练了 240 个 Epoch,每 30 个 Epoch 除以 10。另外,通过大量实验证明公式 12 中的超参数 η 设置为 2,公式 13 中的超参数 α、β、γ 分别设置为 10、1、1,能够使实验结果最优。

如图 3 所示,用 CIFAR-100 数据集在 ResNet32×4 和 ResNet8×4 师生架构下做了大量实验,研究了 α 值的变化对文中方法性能的影响。

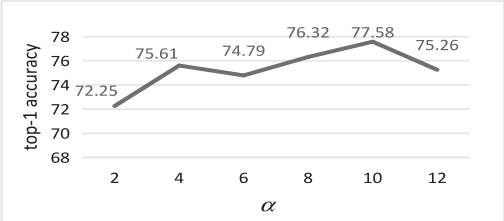


图 3　α 值变化对文中方法性能的影响

表 1　CIFAR-100 数据集同系列师生网络实验结果

蒸馏类型							
	Teacher	ResNet56/% 72.34	ResNet110/% 74.31	ResNet32×4/% 79.42	WRN-40-2/% 75.61	WRN-40-2/% 75.61	VGG13/% 74.64
	Student	ResNet20/% 69.06	ResNet32/% 71.14	ResNet8×4/% 72.5	WRN-16-2/% 73.26	WRN-40-1/% 71.98	VGG8/% 70.36
Logits	KD[14]	70.66	73.08	73.33	74.92	73.54	72.98
	DKD[2]	71.97	74.11	76.32	76.24	74.81	74.68
Features	CRD[26]	71.16	73.48	75.51	75.48	74.14	73.94
	OFD[3]	70.98	73.23	74.95	75.24	74.33	73.95
	FitNet[27]	69.21	71.06	73.5	73.58	72.24	71.02
	RKD[5]	69.61	71.82	71.9	73.35	72.22	71.48
	ReviewKD[4]	71.89	73.89	75.63	76.12	75.09	74.84
Attention	AT[8]	70.55	72.31	73.44	74.08	72.77	71.43
	CAT-KD[9]	71.62	73.62	76.91	75.6	74.82	74.65
	文中方法	72.04	74.19	77.58	76.02	75.14	74.88
	△	0.42	0.57	0.67	0.42	0.32	0.23

表 2 CIFAR-100 数据集不同系列师生网络实验结果

蒸馏类型	Teacher	ResNet32×4/% 79.42	WRN-40-2/% 75.61	ResNet32×4/% 79.42	ResNet50/% 79.34	VGG13/% 74.64
	Student	ShuffleNet-V1/% 70.5	ShuffleNet-V1/% 70.5	ShuffleNet-V2/% 71.82	MobileNet-V2/% 64.6	MobileNet-V2/% 64.6
Logits	KD[14]	74.07	74.83	74.45	67.35	67.37
	DKD[2]	76.45	76.7	77.07	70.35	69.71
Features	CRD[26]	75.11	76.05	75.65	69.11	69.73
	OFD[3]	75.98	75.85	76.82	69.04	69.48
	FitNet[27]	73.59	73.73	73.54	63.16	64.14
	RKD[5]	72.28	72.21	73.21	64.43	64.52
	ReviewKD[4]	77.45	77.14	77.78	69.89	70.37
Attention	AT[8]	71.73	73.32	72.73	58.58	59.4
	CAT-KD[9]	78.26	77.35	78.41	71.36	69.13
	文中方法	78.89	77.8	79.37	72.2	69.79
	△	0.63	0.45	0.96	0.84	0.66

3.3 实验结果

为了验证文中方法的先进性,将文中方法与几种代表性知识蒸馏方法在性能上作比较,如表 1 和表 2 所示。与基准方法 CAT-KD 相比,文中方法在 CIFAR-100 和 ImageNet 数据集上都取得了更好的结果。在 CIFAR-100 数据集上,文中方法在同系列师生网络相较于 CAT-KD 取得了 0.2 ~ 0.7 百分点的提升,在不同系列师生网络中取得了 0.4 ~ 1 百分点的提升。在 ImageNet 数据集上,选取了 ResNet34 和 ResNet18 作为同系列师生网络组,ResNet50 和 MobileNet 作为不同系列师生网络组,如表 3 和表 4 所示,文中方法再次优于其他知识蒸馏方法。

表 3 ImageNet 数据集同系列师生网络实验结果

	Teacher	Student	Features			Logits		Attention		
	ResNet34	ResNet18	OFD[3]	CRD[26]	ReviewKD[4]	KD[14]	DKD[2]	AT[8]	CAT-KD[9]	文中方法
Top-1/%	73.31	69.75	70.81	71.17	71.61	70.66	71.7	70.69	71.26	71.87

表 4 ImageNet 数据集不同系列师生网络实验结果

	Teacher	Student	Features			Logits		Attention		
	ResNet50	MobileNet	OFD[3]	CRD[26]	ReviewKD[4]	KD[14]	DKD[2]	AT[8]	CAT-KD[9]	文中方法
Top-1/%	76.16	68.87	71.25	71.37	72.56	68.58	72.05	69.56	72.24	72.96

3.4 消融实验

在该文提出的基于通道相关性的 CAMs 蒸馏方法中包含了实例内(Intra-KD)CAMs 通道蒸馏和实例间(Inter-KD)CAMs 通道相关性蒸馏。为了进一步了解这两种知识蒸馏对模型性能提升的有效性,通过消融实验探讨了基于通道相关性的 CAMs 蒸馏方法的不同情况,在 CIFAR-100 数据集上使用 ResNet32×4 和 ResNet8×4 分别作为教师和学生模型,并基于 CAT-KD 作对比,如表 5 所示。

表 5 不同方案在 CIFAR-100 数据集上的实验结果

方案	Intra-KD	Inter-KD	Top-1/%	Top-5/%
A	√	√	77.58	95.11
B	√	×	77.32	94.79
C	×	√	75.08	94.06
D	×	×	76.91	94.53

4 结束语

该文提出了一种新的类激活图(CAMs)转移方法,称为基于通道相关性的类注意力知识蒸馏,主要通过学习实例内的 CAMs 通道知识和实例间的 CAMs 通道相关性知识来更好地从教师网络中转移 CAMs。通过转移 CAMs 的方式,学生网络可以学习到教师网络对于识别输入的类别区分区域的能力,从而增强学

生网络在图像分类任务方面的性能。该方法在基于注意力的知识蒸馏方法中都取得了最佳效果,证明了该方法的有效性。

参考文献:

[1] CHENG Y, WANG D, ZHOU P, et al. Model compression and acceleration for deep neural networks: the principles, progress, and challenges[J]. IEEE Signal Processing Magazine, 2018, 35(1): 126-136.

[2] ZHAO B, CUI Q, SONG R, et al. Decoupled knowledge distillation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans: IEEE, 2022: 11953-11962.

[3] HEO B, KIM J, YUN S, et al. A comprehensive overhaul of feature distillation[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE, 2019: 1921-1930.

[4] CHEN P, LIU S, ZHAO H, et al. Distilling knowledge via knowledge review[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville: IEEE, 2021: 5008-5017.

[5] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019: 3967-3976.

[6] TUNG F, MORI G. Similarity-preserving knowledge distillation[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE, 2019: 1365-1374.

[7] YIM J, JOO D, BAE J, et al. A gift from knowledge distillation: fast optimization, network minimization and transfer learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 4133-4141.

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

[9] GUO Z, YAN H, LI H, et al. Class attention transfer based knowledge distillation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Vancouver: IEEE, 2023: 11868-11877.

[10] 乔　虹. 基于注意力机制和多尺度知识蒸馏的图像异常检测[D]. 保定:河北大学,2023.

[11] 陈立玮,周新志. 基于特征自注意力的图像分类知识蒸馏算法[J]. 现代计算机,2023,29(4):49-53.

[12] 高　也. 基于自注意力机制的自知识蒸馏研究[D]. 武汉:华中科技大学,2022.

[13] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 2921-2929.

[14] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv: 1503.02531, 2015.

[15] ZHOU Z, ZHUGE C, GUAN X, et al. Channel distillation: channel-wise attention for knowledge distillation[J]. arXiv: 2006.01683, 2020.

[16] QU Y, DENG W, HU J. H-AT: hybrid attention transfer for knowledge distillation[C]//Pattern recognition and computer vision: third Chinese conference, PRCV 2020. Nanjing: Springer, 2020: 249-260.

[17] SHU C, LIU Y, GAO J, et al. Channel-wise distillation for semantic segmentation[C]//IEEE international conference on computer vision. New York: IEEE, 2020.

[18] YOON D, PARK J, CHO D. Lightweight deep CNN for natural image matting via similarity-preserving knowledge distillation[J]. IEEE Signal Processing Letters, 2020, 27: 2139-2143.

[19] LIU Y, CAO J, LI B, et al. Knowledge distillation via instance relationship graph[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach: IEEE, 2019: 7096-7104.

[20] PASSALIS N, TEFAS A. Learning deep representations with probabilistic knowledge transfer[C]//Proceedings of the European conference on computer vision (ECCV). Zurich: Springer, 2018: 268-284.

[21] PENG B, JIN X, LIU J, et al. Correlation congruence for knowledge distillation[C]//Proceedings of the IEEE/CVF international conference on computer vision. New York, USA: IEEE, 2019: 5007-5016.

[22] YE H J, LU S, ZHAN D C. Distilling cross-task knowledge via relationship matching[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seoul: IEEE, 2020: 12396-12405.

[23] GE S, ZHANG K, LIU H, et al. Look one and more: distilling hybrid order relational knowledge for cross-resolution image recognition[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto: AAAI, 2020: 10845-10852.

[24] ZHU J, TANG S, CHEN D, et al. Complementary relation contrastive distillation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Nashville: IEEE, 2021: 9260-9269.

[25] HUANG T, YOU S, WANG F, et al. Knowledge distillation from a stronger teacher[J]. Advances in Neural Information Processing Systems, 2022, 35: 33716-33727.

[26] TIAN Y, KRISHNAN D, ISOLA P. Contrastive representation distillation[J]. arXiv: 1910.10699, 2019.

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