基于相似一致性的模型自蒸馏方法

万旭，毛莺池，王孜博，刘意，平萍

引用本文

万旭，毛莺池，王孜博，刘意，平萍. 基于相似一致性的模型自蒸馏方法[J]. 计算机科学，2023，50(11): 259-268.

WAN Xu, MAO Yingchi, WANG Zibo, LIU Yi, PING Ping. Similarity and Consistency by Self-distillation Method [J]. Computer Science, 2023, 50(11): 259-268.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000009

基于相似一致性的模型自蒸馏方法

万 旭 毛莺池 王孜博 刘 意 平 萍
水利部水利大数据技术重点实验室 南京 211100
河海大学计算机与信息学院 南京 211100
(211307040041@hhu.edu.cn)

摘 要 针对传统自蒸馏方法存在数据预处理成本高、局部特征检测缺失,以及模型分类精度低的情况,提出了基于相似一致性的模型自蒸馏方法(Similarity and Consistency by Self-Distillation,SCD),提高模型分类精度。首先,对样本图像的不同层进行学习得到特征图,通过特征权值分布获取注意力图。然后,计算 Mini-batch 内样本间注意力图的相似性获得相似一致性知识矩阵,构建基于相似一致性的知识,使得无须对实例数据进行失真处理或提取同一类别的数据来获取额外的实例间知识,避免了大量的数据预处理工作带来的训练成本高和训练复杂的问题。最后,将相似一致性知识矩阵在模型中间层之间单向传递,让浅层次的相似矩阵模仿深层次的相似矩阵,细化低层次的相似性,捕获更加丰富的上下文场景和局部特征,解决局部特征检测缺失问题,实现单阶段单向知识转移的自蒸馏。实验结果表明,采用基于相似一致性的模型自蒸馏方法:在公开数据集 CIFAR100 和 TinyImageNet 上,验证了 SCD 提取的相似一致性知识在模型自蒸馏中的有效性,相较于自注意力蒸馏方法(Self Attention Distillation,SAD)和保持相似性的知识蒸馏方法(Similarity-Preserving Knowledge Distillation,SPKD),分类精度平均提升 1.42%;相较于基于深度监督的自蒸馏方法(Be Your Own Teacher,BYOT)和动态本地集成知识蒸馏方法(On-the-fly Native Ensemble,ONE),分类精度平均提升 1.13%;相较于基于深度神经网络的数据失真引导自蒸馏方法(Data-Distortion Guided Self-Distillation,DDGSD)和基于类间的自蒸馏方法(Class-wise Self-Knowledge Distillation,CS-KD),分类精度平均提升 1.23%。

关键词:知识蒸馏;知识表达;自蒸馏;相似一致性;知识矩阵
中图法分类号 TP311

Similarity and Consistency by Self-distillation Method

WAN Xu,MAO Yingchi,WANG Zibo,LIU Yi and PING Ping
Key Laboratory of Water Big Data Technology of Ministry of Water Resources,Nanjing 211100,China
College of Computer and Information,Hohai University,Nanjing 211100,China

Abstract Due to high data pre-processing costs and missing local features detection in self-distillation methods for models compression,a similarity and consistency by self-distillation(SCD) method is proposed to improve model classification accuracy. Firstly,different layers of the sample images are learned to get the feature maps,and the attention maps are obtained by the distribution of feature weights. Then,the similarity of the attention graph between samples within the mini-batch is calculated to obtain the similar consistency knowledge matrix,and the similar consistency-based knowledge is constructed without distorting the instance data or extracting the same class of data to obtain additional inter-instance knowledge,avoiding a large amount of data preprocessing work. Finally,the similar consistency knowledge matrix is passed unidirectionally between intermediate layers of the model,allowing shallow layers to mimic deep layers and capture richer contextual scenes and local features which can solve the problem of missing local feature detection. Experimental results show that the proposed SCD method can improve the classification accuracy on the public dataset CIFAR100. Compared with the self attention distillation(SAD) method and the similarity-preserving knowledge distillation(SPKD) method,the average improvement is 1.42%. Compared with the be your own teacher (BYOT) method and the on-the-fly native ensemble(ONE) method,the average improvement is 1.13%. Compared with the data-

到稿日期:2022-10-07 返修日期:2023-02-26
基金项目:国家十四五重点研发计划(2022YFC3005401);云南省重点研发计划(202203AA080009,202202AF080003);江苏省科技成果转化项目(BA2021002);江苏省重点研发计划(BE2020729)
This work was supported by the National 14th Five-year Key Research and Development Program of China(2022YFC3005401), Yunnan Province Key Research and Development Program(202203AA080009,202202AF080003),Transformation Program of Scientific and Technological Achievements of Jiangsu Province (BA2021002) and Jiangsu Province Key Research and Development Program(BE2020729).
通信作者:毛莺池(yingchimao@hhu.edu.cn)

distortion guided self-distillation(DDGSD) method and the class-wise self-knowledge distillation(CS-KD) method, the average improvement is 1.23%.

Keywords Knowledge distillation, Knowledge representation, Self-distillation, Similarity consistency, Knowledge matrix

1 引言

近年来,深度神经网络(Deep Neural Networks,DNN)在计算机视觉任务中取得了巨大成功[1-3],如图像分类[4]、目标检测[5]和语义分割[6]。随着 DNN 深度增加,模型通常需要较高的计算成本和推理内存,使用知识蒸馏将教师模型的知识转移到学生模型,以降低计算成本。传统的离线知识蒸馏依赖于强大且受过训练的教师模型[7-9],但训练复杂且高容量的教师模型需要花费大量的时间、计算和存储资源[10]。在线知识蒸馏可以同时更新教师模型和学生模型的参数[11-13],动态建立强大的教师模型,以增强学生网络的学习,降低计算代价,实现高效并行、单阶段端到端的训练。然而,现有在线蒸馏方法通常无法解决模型容量差距问题,当教师模型和学生模型之间容量差异过大时,学生模型很难学习到来自大容量教师模型的知识。因此,本文采用自蒸馏方法,学生模型提取自身知识监督并训练自身学习,解决学生模型和教师模型之间知识学习不充分导致模型分类精度低的问题。自蒸馏方法面临的挑战是如何将学生网络模型自身的知识进行编码和转移,从而提高模型分类精度。现有的自蒸馏方法数据预处理成本高,且深度神经网络每一层关注的信息不同,特别是在层的位置相差较大的情况下,跨越不同层级模仿相似矩阵,忽略中间层特征信息,容易丢失实例间的局部特征信息,使得学习目标单一化和片面化,损害检测局部特征的能力,导致模型分类精度较低。因此,通过自蒸馏方法提高模型的分类精度具有重要的研究意义。

目前大多数自蒸馏方法侧重于如何更好地监督学生模型的自我蒸馏。为了充分挖掘和利用监督学生模型训练的数据,基于数据增强的自蒸馏方法对训练样本进行旋转、剪裁、翻转等操作[14-15],使得学生模型对数据输入不同表示的结果保持不变。在没有分支或其他模型的辅助下,该自蒸馏方法可以高效优化单个学生模型,通过提前进行数据增强的方式,使学生模型具有更多用于泛化的固有表示,显著提高学生模型的分类精度。但该方法需要大量的数据预处理工作,训练复杂且时间、计算成本高。基于架构转换的自蒸馏方法使用代价较小的卷积代替教师模型的标准卷积块进行模型训练,减少内存[16]。虽然基于架构转换的自蒸馏方法可以降低计算代价,减少数据预处理成本,但其在模型训练过程中容易忽略局部特征信息,学习目标缺乏全面性。基于深度监督的自蒸馏方法同样存在局部特征检测缺失的问题,它仅关注单一层次特征区域,导致模型分类精度低。因此,可以在训练过程中逐步提取模型自身的相似一致性知识作为学习目标,并在层间逐层传递一致性知识,引导浅层部分获取深层部分的注意力相关性,充分利用中间层特征信息,细化较低层次的相似性,捕获更加丰富的上下文场景,提升模型的分类精度。

针对自蒸馏方法存在数据预处理成本高和局部特征检测缺失的问题,本文提出了基于相似一致性的模型自蒸馏方法(SCD),将自蒸馏任务转化成学生模型可以结合实际情况和模型自身的相似一致性知识进行适应性调整的任务。首先,为了降低数据预处理成本,构建了基于相似一致性的知识。无须进行数据预处理操作,利用模型自身的相似一致性知识,对样本图像的不同层进行学习得到特征图,基于特征权值分布获得注意力图。然后,计算样本间注意力图的相似性得到相似一致性知识矩阵。最后,在层间逐层传递相似一致性知识矩阵,让浅层相似矩阵近似于深层相似矩阵,细化浅层部分的相似性,解决局部特征检测缺失问题,从而提高模型的分类精度。

本文的主要贡献如下:

(1)基于相似一致性知识构建。利用模型自身的相似一致性知识,降低数据预处理成本。首先,将特征图在特征维度上压平,获取图像的特征权值分布,形成注意力图。然后,网络模型不同层次聚焦注意力图的不同区域,获得不同的特征表达。最后,利用欧拉距离,计算同一层次 Mini-batch 内注意力图的相似性,得到相同层次特征表达间的相似矩阵,层次越靠后,相似矩阵表达的图像间相关关系越清晰。

(2)相似一致性知识传递。使用相似矩阵作为相似一致性的知识进行自蒸馏训练。网络模型深层部分作为浅层部分的教师,监督浅层部分的学习,细化浅层部分的相似性,学习到更准确的信息。浅层部分学习到更准确的特征信息,有助于进一步提升深层部分的特征表达,从而捕获更加全面的局部特征信息。此外,真实标签作为传统监督,用于修正模型的训练方向,从而解决模型局部特征检测缺失问题,提高分类精度。

2 相关工作

传统的自蒸馏方法很大程度上依赖于事先数据增强处理,即对实例数据进行失真处理或提取同一类别的数据,获取额外的实例间知识,大量的数据预处理导致训练时间成本高且训练复杂。基于深度神经网络的数据失真引导自蒸馏方法(DDGSD)使用单一模型提取实例相关信息[14]。首先对样本实例进行随机镜像、旋转等数据增强处理,增加样本丰富度;然后在同一模型下,将样本实例经过不同的失真处理后的数据作为输入,利用 Kullback-Leibler(KL)散度[17]约束用于衡量不同失真样本的后验概率分布匹配。基于类间的自蒸馏方法(CS-KD)在训练期间提取同一标签下不同样本之间的预测分布,以类的方式产生一致的预测,将暗知识(即关于错误预测的知识)规范化,减轻模型过度自信的预测[18]。DDGSD 和 CS-KD 自蒸馏方法从数据角度出发,在不依赖辅助模型的情况下,通过随机的数据增强操作使数据产生差异,在相同训练数据集的不同失真版本之间传递知识[19],优化网络模型,提高模型的分类精度和泛化能力。但是,这类方法数据预处理成本高,且只关注成对的图像,承载的信息有限。更重要的是,它们将常量实例或类别简单定义为正数对,可能导致监督错误。

现有自蒸馏方法研究发现，检测局部特征是模型自蒸馏面临的挑战之一。基于架构转换的自蒸馏方法将教师模型简单转换为学生模型，这种简单转换指保留教师模型的结构，使用成本低的卷积块替换原有标准卷积块，并应用注意力转移（Attention Transfer，AT）将教师模型的注意力图与学生模型的注意力图对齐，生成性能较优的学生模型。动态本地集成知识蒸馏方法（ONE）通过添加辅助分支构造给定目标模型的多分支变体，每个分支与目标模型共享相同的浅层部分，动态集成所有分支构建本地集成教师模型[20]。每个分支不仅受真实标签约束，还受到集成教师模型预测分布的约束。该方法从模型的角度出发，设计辅助网络获取用于自我学习的知识。然而，如果没有合适的辅助网络，浅层部分的特征信息将无法被充分利用，导致局部特征检测缺失，模型缺乏精细知识，从而降低了分类精度。

在神经网络模型中，深层部分比浅层部分包含更多有用和有区分性的特征信息。深层特征是浅层特征的组合，从浅层到深层的特征表示越抽象，就越能表现语义信息或意图，图像表示的类别可以更好地被网络捕获，图像间的相关关系更加明确，更有利于分类。深层分支可以使用基于注意力的方法提取注意力知识，并将注意力知识作为目标，监督浅层分支的训练[21]，还可以为每个分支额外添加辅助分类器，将深层分支分类器提取的特征映射和 logits 作为知识指导浅层分支的训练。基于深度监督的自蒸馏方法（BYOT）利用自提取框架从模型本身提取知识[22]。首先将模型分为几个部分，对每个部分添加相似的辅助网络；然后通过辅助网络得到各个部分的特征知识、logits 层知识和 softmax 层知识；最后用模型深层部分的知识监督浅层部分对应模块的学习，但仍然无法消除辅助网络设计对自蒸馏的影响。自注意力蒸馏方法（SAD）放弃使用辅助网络，直接从经过适当训练的模型中提取注意力图来编码丰富的上下文信息[23]。有价值的上下文信息作为一种“自由”监督的形式，通过在网络模型内进行自上而下和分层的注意力蒸馏，加强表征学习。但是，该方法仅对车道线检测有效。对于车道线检测数据集，仅需关注单一层次特征区域，即车道线存在区域。对于大部分图像分类任务，Block 每层输出关注图像特征的不同区域，SAD 将 Block 每层输出的注意力图作为知识在层间传递，导致学习目标单一化和片面化。BYOT 和 SAD 自蒸馏方法将网络深层部分的知识蒸馏到网络浅层部分，但仍然出现了局部特征丢失的问题。以上方法显著降低了训练的复杂度和时间成本，避免了复杂的多阶段训练过程，但无法解决局部特征检测缺失的问题。

综上所述，用于监督学生模型训练的数据在自蒸馏中发挥了关键作用，构建基于相似一致性的知识可以降低数据预处理成本，更好地监督学生模型自身的训练。将相似一致性知识矩阵在层间单向传递，引导浅层部分获取深层部分的注意力相关性，能够捕获丰富的局部特征信息，避免特征丢失。因此，本文提出基于相似一致性的模型自蒸馏方法（SCD），在训练过程中逐步提取模型自身的相似一致性知识，并将其作为学习目标在层间逐层传递，使浅层部分的相似性被细化，有助于深层部分的表达；使用真实标签来修正模型的训练方向，避免模型在自我学习过程中偏离正确轨迹。

3　总体框架

本文提出的基于相似一致性的模型自蒸馏框架由基于相似一致性知识构建和相似一致性知识传递两部分组成，分别解决数据预处理成本高和局部特征检测缺失的问题。基于相似一致性知识构建部分通过计算 Mini-batch 内实例间的相关关系，得到相似矩阵，降低数据预处理成本。相似一致性知识传递部分则在模型层间传递相似矩阵[24]，细化低层次相似性，捕获丰富的上下文场景和局部特征信息，解决局部特征检测易缺失的问题，提高模型的分类精度。基于相似一致性的模型自蒸馏框架 SCD 如图 1 所示。

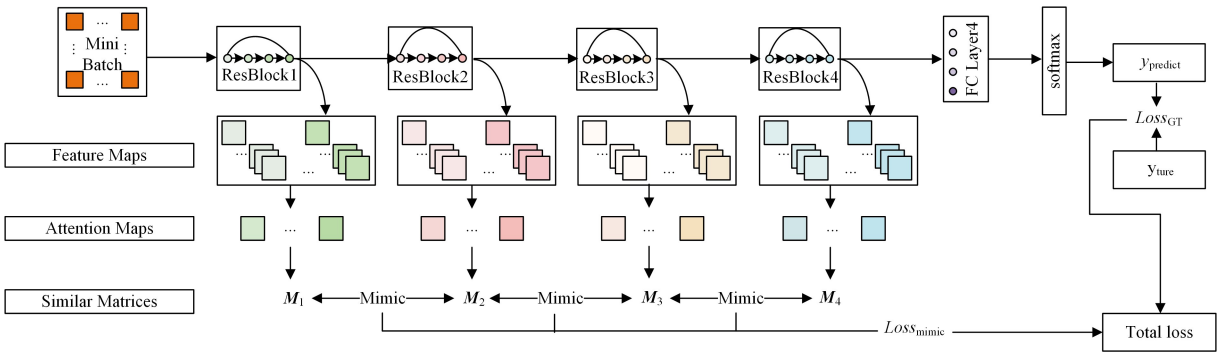


图 1　相似一致性的模型自蒸馏框架图

Fig. 1　Framework of similarity and consistency by self-distillation

综合考虑基于实例关系的相似一致性知识和自蒸馏框架制定本文提出的方法。SCD 框架由基于相似一致性知识构建和相似一致性知识传递两部分组成。其中，基于相似一致性知识构建部分以 ResNet 为例，单个输入在经过每个 ResBlock 卷积操作后，都会生成特征图，将特征图在通道上压平，得到单个输入在经过每个 ResBlock 处理后输出的注意力图。当 Mini-batch 大小为 b 时，每个 ResBlock 会输出 b 个尺寸相同的注意力图。计算注意力图之间的相似性，可以得到大小为 $b\times b$ 的相似矩阵。相似一致性知识传递部分通过使浅层 ResBlock 模仿深层 ResBlock 的相似矩阵，实现相似一致性的模型自蒸馏。此外，真实标签作为传统监督，用于修正模型的训练方向。

4 自蒸馏框架 SCD 描述

4.1 基于相似一致性知识构建

首先将特征图在特征维度上压平，获取图像的特征权值分布，形成注意力图。然后，网络模型不同层次聚焦注意力图的不同区域，获得不同的特征表达。最后利用欧拉距离，计算同一层次 Mini-batch 内注意力图的相似性，得到相同层次特征表达间的相似矩阵。

4.1.1 基于激活的注意力图

注意力图是可以掩码实现的基于特征矩阵的计算方式，可以凝练出有特点的矩阵数据，将图片数据中的关键特征标识出来，经过学习训练得到一组可以作用于原图的权重分布，形成注意力，强调重点区域。通过正确定义神经网络的注意力，可以使用注意力知识训练神经网络，使网络更加关注有效的特征，忽略无效的特征。对最有区别的部分赋予更多的权重，可以更好地捕捉到图像表示的类别，提高网络训练的精度和效率。

注意力可分为强注意力和软注意力。强注意力是随机的预测过程，更强调动态变化，是不可微的，需要通过增强学习来完成训练。软注意力是可微的，可通过训练神经网络计算梯度。

SCD 利用软注意力计算注意力图，考虑神经网络的某一层及其对应的激活张量 $\boldsymbol{A}\in\mathbb{R}^{C\times H\times W}$，其中 $H\times W$ 代表张量的高和宽，C 代表张量的个数，即卷积核的个数。基于激活张量的映射函数，将上述三维激活张量作为输入，输出一个二维空间注意力图，即将三维激活张量在空间维度上压缩为一个平坦的二维向量。

定义 1（软注意力图） 将激活函数的输入张量 $\boldsymbol{A}\in\mathbb{R}^{C\times H\times W}$ 在滤波器个数 C 上压平，即：

$$\mathcal{F}:\mathbb{R}^{C\times H\times W}\rightarrow\mathbb{R}^{H\times W} \tag{1}$$

定义 1 的映射函数将三维特征图映射为二维注意力图，隐含的假设是一个神经元激活的绝对值可以表示激活的重要性。因此，模型考虑张量 $\boldsymbol{A}$ 中元素的绝对值，通过计算这些绝对值在通道维度上的统计量构建一个二维空间注意力图，如图 2 所示。

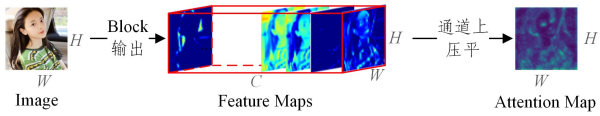


图 2 特征维度上的注意力映射

Fig. 2 Attention mapping on feature dimensions

由定义 1 可知，为了得到张量 $\boldsymbol{A}$ 中元素在通道维度 C 上的绝对值统计量，需要构建注意力映射。可以使用以下 3 种方案：

方案 1 通道 C 的绝对值之和：$F_{\text{sum}}(\boldsymbol{A})=\sum_{i=1}^{C}|\boldsymbol{A}_i|$。

方案 2 通道 C 的绝对值的 p 次幂之和（$p>1$）：$F_{\text{sum}}^{p}(\boldsymbol{A})=\sum_{i=1}^{C}|\boldsymbol{A}_i|^{p}$。

方案 3 通道 C 的绝对值的 p 次幂的最大值（$p>1$）：$F_{\text{max}}^{p}(\boldsymbol{A})=\max_{i=1,C}|\boldsymbol{A}_i|^{p}$。

其中，$\boldsymbol{A}_i=\boldsymbol{A}(i,:,:)$，最大值、幂和绝对值运算是按照元素计算的。

不同的注意力映射函数有不同的属性，如图 3 所示。与 $F_{\text{sum}}(\boldsymbol{A})$ 相比，$F_{\text{sum}}^{p}(\boldsymbol{A})$（$p>1$）更关注空间位置，这些位置对应的神经元具有最高的活跃度。例如：对最有区别的部分赋予更多的权重，即 p 越大，就越关注那些活跃度最高的部分。此外，在所有对应于同一空间位置的神经元激活中，$F_{\text{max}}^{p}(\boldsymbol{A})$ 只考虑其中一个通道来为该空间位置分配权重，与其相反，$F_{\text{sum}}^{p}(\boldsymbol{A})$ 倾向于携带多个高激活神经元的空间位置。

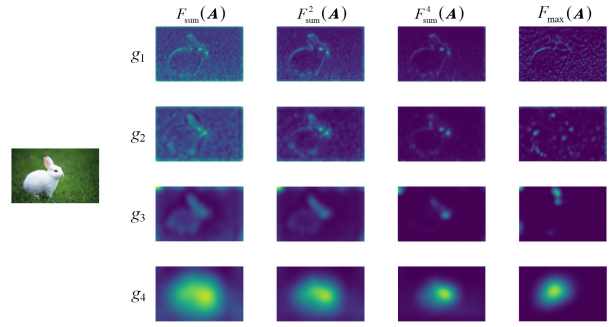


图 3 ResNet-18 网络的激活注意力

Fig. 3 Activation attention of ResNet-18

4.1.2 层次注意力表达

卷积神经网络的权值共享网络结构可以降低网络模型复杂度，减少权值数量。当网络输入是多维图像时，其优势更加明显。图像可以直接作为网络的输入，避免了传统识别算法中复杂的特征提取和数据重建过程。为了识别二维形状，卷积神经网络设计了多层感知器，其网络结构对平移、比例缩放、倾斜或其他形式的变形具有高度不变性。在卷积神经网络中，图像作为层级结构的最底层输入，其信息依次传输到不同的层，每层通过一个数字滤波器获得观测数据的不同特征。

卷积神经网络的不同层次通常聚焦于注意力图的不同部分，如图 4(a)所示。在第一层（g_1），低梯度点的神经元激活水平较高；在中间层（g_2，g_3），最具辨别能力的区域（如眼睛、鼻子和嘴唇周围）的神经元激活水平较高；而在顶层（g_4），它反映完整的物体。这样，在 g_i 中随着 i 的增加，图像的抽象度就越高，存在的猜测就越少，图像表示的类别可以更好地被网络捕获，图像间的相关关系也更加明确。

(a)原始 ResNet-18 网络中不同层次注意力映射图

(b)SCD 训练后 ResNet-18 网络中不同层次注意力映射图

图 4 自蒸馏前后 ResNet-18 网络中不同层次注意力映射图

Fig. 4 Attention maps of different layers in ResNet-18 before and after self-distillation

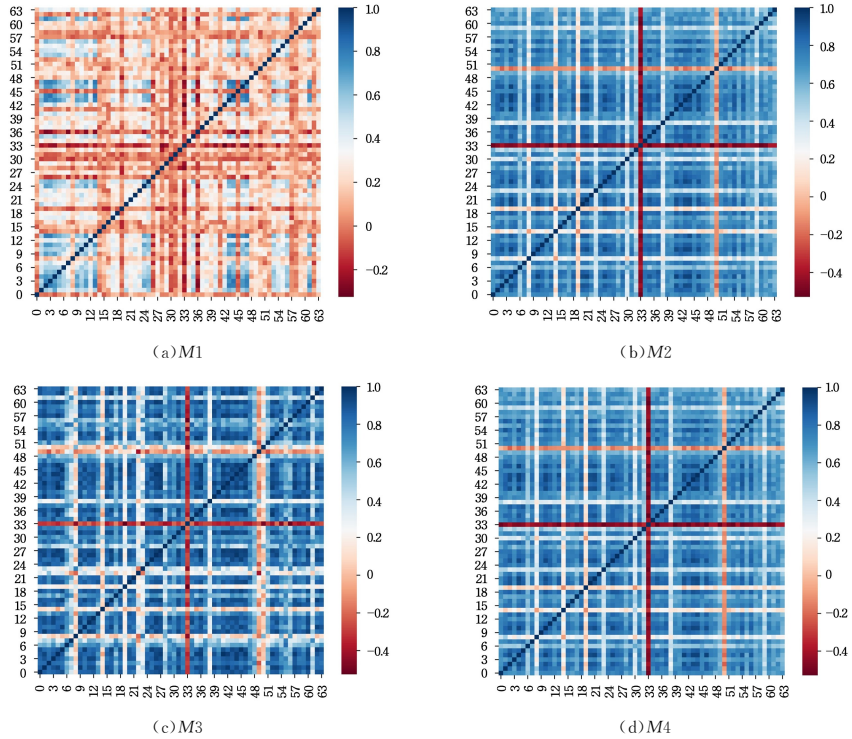
原始网络模型在由基于相似一致性的模型自蒸馏方法训练后，网络中间层注意图如图 4(b)所示。在眼睛、鼻子、耳朵和

嘴唇周围会有更高的激活，此外还捕捉到了更多的局部特征信息（如下巴和发际线周围），顶层的激活则对应于整个图像。

4.1.3　相似矩阵层次表达

在神经网络的训练过程中，语义上相似的输入倾向于诱发相似的激活模式，如图 5 所示。可以观察到，激活模式在同一对象类别中基本上是一致的，而在不同类别中是不同的。激活模式中的相关性是否对可转移的有用知识进行编码？假设两种输入在神经网络中产生了高度相似的激活模式，那么引导位置靠前的 ResBlock 向位置靠后的 ResBlock 中高度相似激活的方向发展就是有益的。相反，如果两种输入在位置靠后的 ResBlock 中产生了不同的激活模式，则期望这些输入在位置靠前的 ResBlock 上也产生不同的激活模式。

图 5 给出了 CIFAR10 测试集中，ResNet-18 在 Mini-batch 上根据不同深度位置 ResBlock 得到的相似矩阵（图 1 所示 $M1$，$M2$，$M3$，$M4$）。横纵坐标表示 Mini-batch 内图像的编号，可以看到，ResBlock 位置越靠后，特征抽象程度越高，图像类别属性越明显，Mini-batch 内图像间的相关关系越明确。



（a）$M1$　（b）$M2$　（c）$M3$　（d）$M4$

图 5　ResBlock 基于注意力图的相似矩阵

Fig. 5　Similarity matrix based on attention maps in ResBlock

将大小为 b 的 Mini-batch 输入网络模型 ResNet 中，经过第 i 个 ResBlock 卷积操作后生成的激活映射（即特征图）为：$\boldsymbol{A}^{(i)} \in \mathbb{R}^{b\times c_i\times h_i\times w_i}$，$1\leqslant i\leqslant 4$。其中 b 是 batch size 大小，c_i 是第 i 个 ResBlock 滤波器个数，h_i 和 w_i 分别表示第 i 个 ResBlock 输出特征图的高和宽。

根据 4.1.1 节提出的注意力图，可将特征图在滤波器个数维度上压平，即 $F:\mathbb{R}\rightarrow\mathbb{R}^{b\times h_i\times w_i}$，得到 b 个大小为 $h_i\times w_i$ 的注意力图。计算 b 个注意力图之间的相关性，得到注意力图之间的相关性矩阵 $\boldsymbol{M}_i\in\mathbb{R}^{b\times b}$，如图 6 所示。$\boldsymbol{M}_i$ 中的元素 $M_i^{p,q}$ 为：

$$M_i^{p,q}=x_{pq}=\frac{\|x_p^A-x_q^A\|_2}{\|x_p^A-x_q^A\|} \tag{2}$$

其中，x_p^A 和 x_q^A 分别代表 Mini-batch 内的第 p 个和第 q 个图像的注意力图。

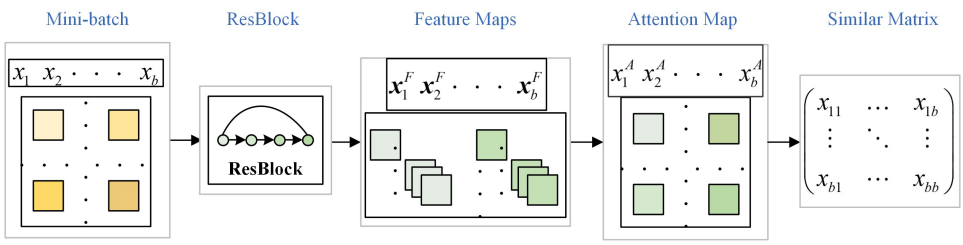


图 6　相似矩阵的计算

Fig. 6　Calculation of similarity matrix

4.2　相似一致性知识传递

SCD 采用自上而下的注意力相似性蒸馏，增强表示学习的过程。由于相似矩阵来自模型本身，因此，本文提出的相似一致性蒸馏无需任何外部监督或附加标签，只需引入保持相似一致性的自我知识蒸馏，使用每个输入 Mini-batch 中的成对激活相似性来监督训练。保持相似一致性的自我知识蒸馏

是根据激活函数定义的，而不是基于传统知识蒸馏中的类别。在 Fitnets、基于流动的知识蒸馏和注意力转移中，激活函数也被用于定义蒸馏损失。但是，这些蒸馏方法鼓励学生模型模仿教师模型表现空间的不同方面。保持相似一致性的模型自蒸馏方法保持了 batch 内输入样本间的激活相似性，其行为不会因教师模型表示空间的旋转而改变。学生模型也无需表达教师模型的表示空间，仅需在学生空间中保留教师空间的相似性。在模型自蒸馏过程中，将位置靠后的 ResBlock 当作教师，将位置靠前的 ResBlock 当作学生。在层间逐层传递相似一致性知识矩阵，让浅层相似矩阵近似于深层相似矩阵。

为引导浅层部分获取深层部分的注意力相关性，使用较深的分支提取知识，并将提取的知识转移到浅层分支，利用深层的信息来反哺优化浅层，细化浅层部分的相似性，以此学习到更准确的信息。浅层部分学习到的更准确的特征表达，又能进一步帮助提升深层部分的特征表达。如图 4 所示，浅层部分的特征相似性提高，有助于包含更多有用信息的深层部分的特征表达，使模型捕获到更加全面的局部特征信息，加强网络整体的特征表达能力，明确图像间的相关关系，捕获更加丰富的上下文场景，从而提高模型分类精度。

给定 M 个类中的 N 个样本 $X=\{x_i\}_{i=1}^{N}$，对应的标签集合为 $Y=\{y_i\}_{i=1}^{M}$，$y_i\in\{1,2,\cdots,M\}$。网络模型中多个 ResBlock 的输出相似矩阵表示为 $\{\boldsymbol{M}_i\}_{i=1}^{4}$。

在模型自蒸馏过程中，为了引导 Mini-batch 较低层次 Block 获取高层次 Block 的注意力相关性，我们让浅层相似矩阵近似于深层相似矩阵，表示相似一致性知识监督。得到的相似一致性 SCD 损失($\text{Loss}_{\text{mimic}}$)如式(3)所示：

$$L_{\text{mimic}}=\sum_{i=1}^{3}\left\|\frac{\boldsymbol{M}_{i+1}}{\|\boldsymbol{M}_{i+1}\|_2}-\frac{\boldsymbol{M}_i}{\|\boldsymbol{M}_i\|_2}\right\|_p \tag{3}$$

其中，$\boldsymbol{M}_i$ 是第 i 个 ResBlock 输出得到的相似矩阵；p 取值为 2，使用欧拉距离度量不同层次相似矩阵的差异程度。

除了相似一致性知识监督外，还需要依赖于真实标签的监督(Loss_{GT})，如式(4)所示：

$$L(W_R,x)=CrossEntropy(y_{\text{predict}},y) \tag{4}$$

其中，W_R 是模型参数，y_{predict} 是输入 x 的 softmax 层的输出，y 是 x 的真实标签。

综上所述，整个神经网络的损失函数(Total Loss)由自注意力损失和真实标签损失组成，可表示为：

$$L_{\text{KD}}=\alpha L(W_R,x)+\beta\sum_{i=1}^{3}\left\|\frac{\boldsymbol{M}_i}{\|\boldsymbol{M}_i\|_2}-\frac{\boldsymbol{M}_{i+1}}{\|\boldsymbol{M}_{i+1}\|_2}\right\|_p \tag{5}$$

其中，α 和 β 为超参数。

SCD 路径可以推广到其他密集连接。例如，可以将路径块表示为：$ResBlock1\xrightarrow{\text{mimic}}ResBlock3$，$ResBlock2\xrightarrow{\text{mimic}}ResBlock4$ 等。一般情况下，深度为 d 层的网络可能的 SCD 路径数为 $d(d-1)/2$，我们将在实验中评估不同路径的可能性。

基于相似一致性的模型自蒸馏方法如算法 1 所示。算法 1 总结了模型的训练和部署细节，目标网络模型自身进行协作训练。知识蒸馏在每一个 batch 内进行，并贯穿整个训练过程。由于只对一个网络的 ResBlock 进行训练，因此只需对 ResBlock 进行随机梯度下降(Stochastic Gradient Descent，SGD)，训练整个网络直到收敛。当模型训练完成后，在测试时不需要计算测试数据集的相似矩阵，即可获得用于部署的原始网络体系结构。因此，SCD 方法不会增加测试时间成本。

算法 1 基于相似一致性的模型自蒸馏 SCD

输入：训练数据集 D_{train}；训练轮次 τ；batch size 大小 b；

输出：目标 CNN 模型 θ^t

1. / * Training * /
2. 初始化：t=1；
3. while $t\leqslant\tau$ do
4. $\mathcal{F}:\mathbb{R}^{C\times H\times W}\rightarrow\mathbb{R}^{H\times W}$ //将 ResBlock 输出的特征图在通道上压平得到注意力图
5. $M_i^{p,q}=x_{pq}=\dfrac{\|x_p^A-x_p^A\|_2}{\|x_p^A-x_p^A\|}$ //计算 batch 内数据间注意力图的相似性
6. $L_{\text{mimic}}=\sum_{i=1}^{3}\left\|\dfrac{\boldsymbol{M}_{i+1}}{\|\boldsymbol{M}_{i+1}\|_2}-\dfrac{\boldsymbol{M}_i}{\|\boldsymbol{M}_i\|_2}\right\|_p$ //使 ResBlock_i 逼近 ResBlock_{i+1} 的相似矩阵
7. $L(W_R,x)=\text{CrossEntropy}(y_{\text{predict}},y)$ //模型的真实标签监督
8. $L_{\text{KD}}=\alpha L(W_R,x)+\beta\sum_{i=1}^{3}\left\|\dfrac{\boldsymbol{M}_i}{\|\boldsymbol{M}_i\|_2}-\dfrac{\boldsymbol{M}_{i+1}}{\|\boldsymbol{M}_{i+1}\|_2}\right\|_p$ //最终损失函数
9. update the model parameters by SGD
10. t=t+1
11. end
12. / * Testing * /

输入：测试数据集 D_{test}

输出：预测精度

算法 1 第 1－2 行表示算法的初始化；第 3－9 行迭代更新模型参数，其中第 4－5 行计算 batch size 内输入数据之间的相似矩阵，第 6－8 行表示算法的损失函数，第 9 行表示算法通过随机梯度下降进行参数更新。最终，当损失趋于零且稳定时，基于相似一致性的模型自蒸馏方法训练完成。在测试时，不需要计算 batch size 内测试数据之间的相似矩阵，即可获得用于部署的原始网络体系结构，不会增加测试时间成本。

本文利用模型自身的相似一致性知识，构建相似一致性知识矩阵，并在层间单向逐层传递，引导浅层部分获取图像间相关关系更明确的深层部分的注意力相关性，让浅层相似矩阵近似于深层相似矩阵，细化浅层部分的相似性。本文方法降低了数据预处理成本，且充分利用了中间层特征信息，捕获了更加全面的局部特征信息，加强了网络整体的特征表达能力，提高了模型分类精度。

5 实验验证

5.1 实验准备

5.1.1 数据集和评价指标

在实验中使用 5 个多类分类基准数据集，如表 1 所列。

(1)CIFAR10[25]：自然图像数据集，包含从 10 个类中抽取的 50000/10000 个训练/测试样本，总共 60000 张图，每个类有 6000 张 32×32 像素的图像。

(2)SVHN[26]：街景房屋编号(SVHN)数据集，由 73257/26032 个标准训练/测试图像和另外一组 531131 张训练图像组成，每张图片包含若干数字的门牌号，标签从 0－10，且

所有数字都已调整为 32×32 像素的固定分辨率。

(3)FashionMNIST[27]：服装数据集，包含从 10 个类别抽取的 60000/10000 个训练/测试样本，总共 70000 张图，每个类有 7000 张 28×28 像素的图像。

(4)CIFAR100[25]：一个类似 CIFAR10 的数据集，也包含 50000/10000 个训练/测试图像，但覆盖 100 个类，每个类有 600 张图片。

(5)TinyImageNet[28]：ImageNet 数据集的修改子集，包含 100000/10000 个训练/测试图像，覆盖 200 个类，图像为 64×64 像素。

表 1　检测分类数据集

Table 1　Classification data set detection

数据集	类别	训练样本	测试样本
CIFAR10	10	50000	10000
SVHN	10	73257	26032
FasionMNIST	10	60000	10000
CIFAR100	100	50000	10000
TinyImageNet	200	100000	10000

本文实验采用评价指标为常见的 Top-n(n=1,5)分类准确率/错误率和每轮平均分类训练损失。

5.1.2　网络模型和基准方法

为了验证 SCD 在多个数据集上的有效性，本文研究了网络模型 VGG19 和原始 ResNet 网络模型及其变体(包括 ResNet，DenseNet，Wide-ResNet，ResNetXt)。此外，将 SCD 与 7 种基准方法进行比较，具体如下：

(1)原始分类器(Baseline)：该分类器仅使用基于交叉熵的损失。

(2)BYOT[22]：通过真实标签和网络自身信号(预测 logit 和特征图)训练辅助分类器。

(3)ONE[20]：利用附加分支的综合预测作为软标签。

(4)DDGSD[14]：通过单个实例生成不同版本(镜像或旋转)，并对不同版本实例进行一致预测。

(5)CS-KD[18]：通过对与软标签相同的类中的其他实例的预测，强制相同类进行一致预测。

(6)SPKD[29]：通过实例在 Mini-batch 内的相似性，进行相似性一致的预测。

(7)SAD[23]：通过网络本身的分层注意力蒸馏来进行车道线检测。

5.1.3　实验参数设置

在 Pytorch 中实现网络和模型训练过程。根据数据集应用相同的训练设置，调优网络模型和基准方法的超参数。所有分类实验均使用带有动量的随机梯度下降(SGD)进行参数学习和更新操作，初始学习率设置为 0.1，动量设置为 0.9，权重衰减为 5×10^{-4}。超参数 α 设置为 0.5，β 设置为 2。batch size 设置为 64，Epoch 设置为 100，在 50%的训练时学习率从 0.1 降至 0.01，在 75%时降至 0.001。由于单次实验具有随机性，采用 5 次独立实验的平均分类错误率作为评判依据。

5.2　实验结果与分析

5.2.1　准确性分析

(1)与网络模型比较

因为数据集类别数目大小不同，分类难易程度也不同。我们用 Top-1 错误率作为类别数目较小数据集(CIFAR10，SVHN，FashionMNIST)的评价指标；对于类别数目较大的数据集(CIFAR100)，使用 Top-1 错误率和 Top-5 错误率作为评价指标。

为了验证所提出的 SCD 方法可以提高网络模型分类的准确率，将其分别应用于网络模型 DenseNet-121，ResNet-18，ResNet-32 和 ResNeXt-50(32×4d)上。原始学习方法和在原始学习方法上运用本文提出的 SCD 方法训练的 4 种不同容量的最新网络模型在数据集 CIFAR10，SVHN，FashionMNIST 上的分类 Top-1 错误率结果如表 2 所列。

表 2　SCD 方法在数据集 CIFAR10，SVHN，FashionMNIST 上的分类 Top-1 错误率

Table 2　Top-1 error rate of SCD method for classification on datasets CIFAR10，SVHN，FashionMNIST

Method	CIFAR10	SVHN	FashionMNIST	Params
DenseNet-121	4.71	2.31	4.47	7.98×10^6
DenseNet-121+SCD	**4.02±0.03**	**2.01±0.02**	**3.84±0.05**	7.98×10^6
ResNet-18	7.07	8.04	13.29	11.69×10^6
ResNet-18+SCD	**4.08±0.05**	**7.79±0.06**	**11.81±0.07**	11.69×10^6
ResNet-32	6.92	2.11	10.31	21.80×10^6
ResNet-32+SCD	**4.62±0.07**	**1.78±0.05**	**6.56±0.05**	21.80×10^6
ResNeXt-50(32×4d)	3.21	2.14	4.01	25.03×10^6
ResNeXt-50(32×4d)+SCD	**2.71±0.02**	**2.10±0.03**	**3.81±0.05**	25.03×10^6

从表 2 可以看出，在数据集 CIFAR10，SVHN，FashionMNIST 上，运用本文提出的 SCD 方法后，DenseNet-121 的分类精度分别提升 0.69%，0.3%，0.63%，平均提升 0.54%；ResNet-18 的分类精度分别提升 2.99%，0.25%，1.48%，平均提升 1.57%；ResNet-32 的分类精度分别提升 2.3%，0.33%，3.57%，平均提升 2.06%；ResNeXt-50(32×4d)的分类精度分别提升 0.5%，0.04%，0.2%，平均提升 0.25%。4 种网络模型都受益于 SCD 方法，这表明，SCD 在从在线教师模型到目标模型的在线知识提取方面具有普遍的优势。这是因为 SCD 利用 Mini-batch 内样本的相似一致性，增强了表示学习的过程。

DenseNet 网络模型虽然参数量少，但其在各个数据集上的性能都比较优越，这是由于与 ResNet 相比，DenseNet 能够解决在小数据集上的过拟合问题，具有较好的抗过拟合性。运用 SCD 方法后，DenseNet 的性能更加优越。

将 SCD 分别应用在网络模型 ResNet-18，ResNet-110 和 Wide-ResNet28×10 上，原始学习方法和在原始学习方法上运用本文提出的 SCD 方法训练的 3 种不同容量的最新网络模型的 Top-1 和 Top-5 错误率在数据集 CIFAR100 上的分类 Top-1 错误率、Top-5 错误率结果如表 3 所列。

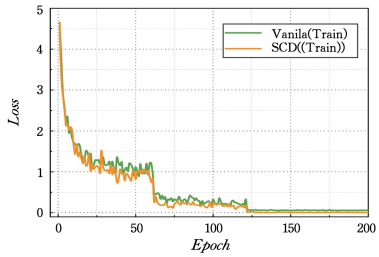
表 3 SCD 方法在数据集 CIFAR100 上的分类 Top-1 错误率、Top-5 错误率

Table 3 Top-1 error rate, Top-5 error rate of SCD method for classification on dataset CIFAR100

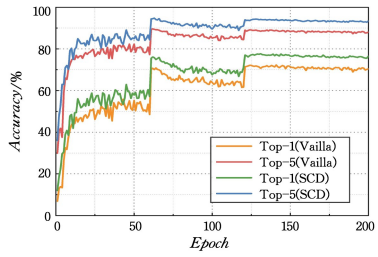
Method	Top-1	Top-5
ResNet-18	33.40	9.43
ResBet-18+SCD	**32.58±0.04**	**8.97±0.12**
ResNet-110	22.02	6.42
ResNet-110+SCD	**20.97±0.11**	**5.13±0.06**
Wide-ResNet28×10	22.70	5.60
Wide-ResNet28×10+SCD	**22.12±0.07**	**5.05±0.12**

从表 3 可以看出，在数据集 CIFAR100 上，运用本文提出的 SCD 方法后，ResNet-18 的 Top-1 和 Top-5 正确率分别提升 0.82%和 0.46%；ResNet-110 的 Top-1 和 Top-5 正确率分别提升 1.05%和 1.29%；Wide-ResNet28×10 的 Top-1 和 Top-5 正确率分别提升 0.58%和 0.55%。说明本文提出的 SCD 方法仍然能产生有效的训练和更一般化的模型，SCD 方法可以普遍应用于分类类别数目较大的图像分类设置。

图 7(a)给出了数据集 CIFAR100 在 Wide-ResNet28×10 和 Wide-ResNet28×10+SCD 上的训练损失，可以看到，运用 SCD 方法的自蒸馏相较于原始学习方法的训练损失更低，训练过程更加平稳。这表明，基于相似一致性的模型自蒸馏方法 SCD 对图像分类具有积极的正则化效应。图 7(b)给出了 Wide-ResNet28×10 和 Wide-ResNet28×10+SCD 在数据集 CIFAR100 上的 Top-1 和 Top-5 的测试准确率，可以看出 Wide-ResNet28×10+SCD 的 Top-1 和 Top-5 的测试准确率均优于原始学习方法。



(a)每轮平均训练损失



(b)测试准确率

图 7 SCD 训练前后网络模型 Wide-ResNet28×10 的训练损失和测试准确率

Fig. 7 Training loss and test accuracy of Wide-ResNet28×10 before and after SCD training

(2)与基准方法比较

将 SCD 及 6 种具有代表性的(自)蒸馏方法分别应用于原始分类器(Baseline)上的分类 Top-1 错误率比较结果如表 4 所列，—表示 ResNet-32 作为教师模型的结果。

表 4 SCD 与基准方法在数据集 CIFAR100 和 TinyImageNet 上的分类 Top-1 错误率

Table 4 Top-1 classification error rate of SCD and baseline methods on CIFAR100 and TinyImageNet datasets

Baseline	Method	Dataset	
		CIFAR100	TinyImageNet
WRN-16-2	WRN-16-2	29.58±0.08	48.95±0.20
	WRN-16-2+DDGSD	28.04±0.05	47.93±0.24
	WRN-16-2+CS-KD	28.21±0.68	46.62±0.18
	WRN-16-2+ONE	26.99±0.23	45.90±0.20
	WRN-16-2+BYOT	29.78±0.26	44.67±0.03
	WRN-16-2+SAD	29.69±0.45	47.74±0.39
	WRN-16-2+SPKD	27.01±0.22	46.09±0.30
	WRN-16-2+SCD	**25.31±0.09**	**43.92±0.33**
ResNet-32	ResNet-32	24.20±0.70	46.40±0.89
	ResNet-32+DDGSD	22.38±0.47	44.25±0.24
	ResNet-32+CS-KD	21.82±0.05	43.53±0.10
	ResNet-32+ONE	22.31±0.54	45.74±0.39
	ResNet-32+BYOT	22.32±0.07	44.31±0.30
	ResNet-32+SAD	23.35±0.32	45.05±0.06
	ResNet-32+SPKD	—	43.59±0.44
	ResNet-32+SCD	**21.67±0.11**	**42.54±0.04**
VGG19	VGG19	31.53±0.55	53.47±0.39
	VGG19+DDGSD	30.99±0.24	53.07±0.14
	VGG19+CS-KD	31.24±0.29	52.33±0.09
	VGG19+ONE	31.17±0.85	52.60±0.20
	VGG19+BYOT	30.94±0.11	52.77±0.56
	VGG19+SAD	31.32±0.25	52.19±0.35
	VGG19+SPKD	31.02±0.15	53.26±0.46
	VGG19+SCD	**30.67±0.12**	**52.15±0.18**
DenseNet-121	DenseNet-121	22.27±0.10	38.56±1.58
	DenseNet-121+DDGSD	22.08±0.50	38.44±0.12
	DenseNet-121+CS-KD	21.99±0.49	37.96±0.09
	DenseNet-121+ONE	21.81±0.28	38.10±0.17
	DenseNet-121+BYOT	21.97±0.83	37.98±0.52
	DenseNet-121+SAD	22.11±0.03	38.21±0.15
	DenseNet-121+SPKD	21.88±0.25	38.37±0.50
	DenseNet-121+SCD	**21.75±0.37**	**37.61±0.38**

使用 WRN-16-2，ResNet-32，VGG19 和 DenseNet-121 作为目标模型在数据集 CIFAR100 和 TinyImageNet 上评估不同自蒸馏方法的性能，从表 4 可以看出，大多数自蒸馏方法都能提升原始分类器的性能，其中 SCD 表现出了比其他自蒸馏方法更好的性能。这是由于 ONE 和 BYOT 依赖辅助网络的设计，如果没有合适的辅助分支，那么其浅层部分特征信息就无法被充分利用，模型缺乏精细知识，分类精度不佳。SAD 仅适用于单一特征实例，对大部分图像分类任务而言，学习目标片面化，易丢失局部特征信息，导致模型分类精度低。而 SCD 关注 Mini-batch 内图像的相似一致性，在网络各层之间转移基于相似一致性的知识进行自我学习，使得较低层次的相似性被细化。较低层次学习到的知识对于较深层次的表达是有利的，可以捕获更加丰富的上下文场景，提高模型分类精度。SCD 比 SPKD 更注重自我提升，其不需要预训练教师模型，更加关注自身学习进度，整个学习训练过程循序渐进，避免了教师模型和学生模型容量差距过大导致学生模型无法充分利用教师模型的知识的问题。除此以外，SCD 利用自身的相似一致性知识，解决了 DDGSD 和 CS-KD 带来的数据预处理成本高的问题。

5.2.2 注意力图映射方案分析

如 4.1.1 节所述，为得到张量 $\boldsymbol{A}$ 中元素在通道维度 C 上

的绝对值统计量，需要构建注意力映射。在这里，我们可以使用如下 4 种方案：

方案 1　通道 C 的绝对值之和：$F_{\text{sum}}(\boldsymbol{A})=\sum_{i=1}^{C}|A_i|$，即 g.mean(1)。

方案 2　通道 C 的绝对值的 p 次幂之和($p>1$)：$F_{\text{sum}}^{p}(\boldsymbol{A})=\sum_{i=1}^{C}|A_i|^p$($p=2$)，即 g.pow(2).mean(1)。

方案 3　通道 C 的绝对值的 p 次幂之和($p>1$)：$F_{\text{sum}}^{p}(\boldsymbol{A})=\sum_{i=1}^{C}|A_i|^p$($p=4$)，即 g.pow(4).mean(1)。

方案 4　通道 C 的绝对值的 p 次幂的最大值($p>1$)：$F_{\text{sum}}^{p}(\boldsymbol{A})=\max_{i=1,C}|A_i|^p$，即 g.max()。

在数据集 CIFAR10 上，使用网络模型 ResNet-18，不同注意力映射方案的错误率结果如表 5 所列。

表 5　注意力映射方案

Table 5　Attention Mapping Scheme

Attention Scheme	ResNet-18
g.mean(1)	3.74±0.06
g.pow(2).mean(1)	**2.08±0.05**
g.pow(4).mean(1)	2.17±0.07
g.max()	3.88±0.11

从图 8 可以看出，当注意力映射方案为 g.pow(2).mean(1)，即 $F_{\text{sum}}^{p}(\boldsymbol{A})=\sum_{i=1}^{C}|A_i|^p$($p=2$)时模型效果最好。因为 $F_{\text{sum}}^{p}(\boldsymbol{A})$倾向于携带多个高激活神经元的空间位置，关注更多重要元素，保留了更多利于分类的有效信息。

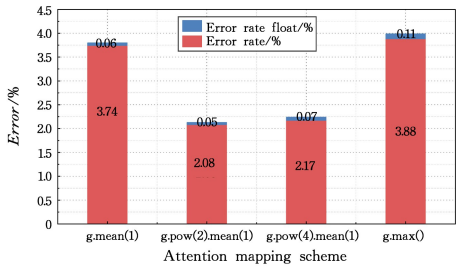


图 8　不同注意力映射方案错误率

Fig. 8　Error rate of different attention mapping schemes

5.2.3　自蒸馏路径选择分析

自蒸馏过程中的路径选择如下所示：

$$Path1=P_{12}+P_{23}+P_{34}$$

$$Path2=P_{12}+P_{24}+P_{34}$$

$$Path3=P_{13}+P_{23}+P_{34}$$

$$Path4=P_{13}+P_{24}+P_{34}$$

$$Path5=P_{14}+P_{23}+P_{34}$$

$$Path6=P_{14}+P_{24}+P_{34}$$

其中，P_{ij} 表示 $ResBlock\ i \xrightarrow{\text{mimic}} ResBlock\ j$。在数据集 CIFAR10 上，使用网络模型 ResNet-32，自蒸馏不同路径下的分类准确率如表 6 所列。

表 6　蒸馏不同路径下的分类准确率

Table 6　Classification accuracy with different paths of self-distillation

Path	Path1	Path2	Path3	Path4	Path5	Path6
Accuracy	98.22	97.82	96.01	94.35	97.55	90.41

从图 9 可以看出，路径 1 的分类准确率和自蒸馏精度均最高。这是由于自我学习是一个循序渐进的过程，在语义层面上，相邻层的相似矩阵比非相邻层的相似矩阵更接近。路径 4 和路径 6 准确率较差，原因在于深度神经网络每一层关注的信息不相同，特别是在层的位置相差较大的情况下，跨越不同层级模仿相似矩阵，容易丢失局部特征信息。

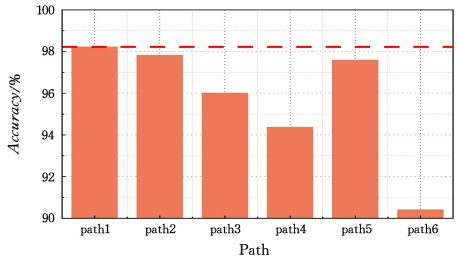


图 9　不同路径下的分类准确率

Fig. 9　Classification accuracy with different paths

5.3　实验小结

通过在多个数据集上进行对比实验，验证了 SCD 能有效提升模型的分类性能。在公开数据集 CIFAR100 和 Tiny-ImageNet 上，相较于 SAD 和 SPKD，SCD 分类精度平均提升 1.42%；相较于 BYOT 和 ONE，分类精度平均提升 1.13%；相较于 DDGSD 和 CS-KD，分类精度平均提升 1.23%。同时，通道压平的最优方法是 $F_{\text{sum}}^{p}(\boldsymbol{A})=\sum_{i=1}^{C}|A_i|^p$($p=2$)；自蒸馏的最优路径是 $Path=P_{12}+P_{23}+P_{34}$。

结束语　本文分析了现有知识蒸馏领域中的自蒸馏方法，针对数据预处理成本高、局部特征检测缺失导致模型分类精度低的问题，本文提出了基于相似一致性的模型自蒸馏方法 SCD。首先通过样本图像的特征权值分布得到注意力图；然后网络模型不同层次聚焦注意力图的不同区域，获得不同的特征表达；接着计算同一层次 Mini-batch 内注意力图的相似性，得到相同层次特征表达间的相似矩阵；最后将相似矩阵作为相似一致性的知识进行自蒸馏训练。实验结果表明，本文提出的模型自蒸馏方法能够有效降低数据预处理成本并捕获丰富的局部特征信息，同时效果优于当前所有网络模型和基准方法，使得自蒸馏的模型分类性能显著提高成为可能。

参 考 文 献

[1] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 770-778.

[2] LI W, ZHU X, GONG S. Person Re-Identification by Deep Joint Learning of Multi-Loss Classification [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: Morgan Kaufmann, 2017: 2194-2200.

[3] LAN X, ZHU X, GONG S. Person Search by Multi-Scale Matching[C]// Proceedings of European Conference on Computer Vision. Munich: Springer Verlag, 2018: 536-552.

[4] XIE G S, ZHANG Z, LIU L, et al. SRSC: Selective, Robust, and Supervised Constrained Feature Representation for Image Classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(10): 4290-4302.

[5] CAI Q, PAN Y, WANG Y, et al. Learning a Unified Sample

Weighting Network for Object Detection[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Online: IEEE Press, 2020: 14173-14182.
[6] LIU Y, CHEN K, LIU C, et al. Structured Knowledge Distillation for Semantic Segmentation[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE Press, 2019: 2604-2613.
[7] PASSALIS N, TZELEPI M, TEFAS A. Heterogeneous Knowledge Distillation Using Information Flow Modeling[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Greece: IEEE Press, 2020: 2339-2348.
[8] ZHAO L, PENG X, CHEN Y, et al. Knowledge as Priors: Cross-Modal Knowledge Generalization for Datasets Without Superior Knowledge[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Online: IEEE Press, 2020: 6528-6537.
[9] XU G, LIU Z, LI X, et al. Knowledge Distillation Meets Self-Supervision[C]// Proceedings of European Conference on Computer Vision. Glasgow: Springer, 2020: 588-604.
[10] ANIL R, PEREYRA G, PASSOS A, et al. Large Scale Distributed Neural Network Training Through Online Distillation[C]// International Conference on Learning Representations. 2018: 1-12.
[11] CHEN D, MEI J P, WANG C, et al. Online Knowledge Distillation with Diverse Peers[C]// Proceedings of AAAI Conference on Artificial Intelligence. New York: AAAI press, 2020: 3430-3437.
[12] GUO Q, WANG X, WU Y, et al. Online Knowledge Distillation via Collaborative Learning[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Online: IEEE Press, 2020: 11017-11026.
[13] WU G, GONG S. Peer Collaborative Learning for Online Knowledge Distillation[C]// Proceedings of AAAI Conference on Artificial Intelligence. 2021.
[14] XU T B, LIU C L. Data-Distortion Guided Self-Distillation for Deep Neural Networks[C]// Proceedings of AAAI Conference on Artificial Intelligence. Honolulu: AAAI press, 2019: 5565-5572.
[15] LEE H, HWANG S J, SHIN J. Rethinking Data Augmentation: Self-Supervision and Self-Distillation[J]. arXiv: 1910. 05872, 2019.
[16] CROWLEY E J, GRAY G, STORKEY A J. Moonshine: Distilling with Cheap Convolutions[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems. 2018: 2888-2898.
[17] BARZ B, RODNER E, GARCIA Y G, et al. Detecting Regions of Maximal Divergence for Spatio-Temporal Anomaly Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(5): 1088-1101.
[18] YUN S, PARK J, LEE K, et al. Regularizing Class-Wise Predictions via Self-Knowledge Distillation[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Online: IEEE Press, 2020: 13873-13882.
[19] XU T B, LIU C L. Deep Neural Network Self-Distillation Exploiting Data Representation Invariance[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 33(1): 257-269.
[20] LAN X, ZHU X, GONG S. Knowledge Distillation by On-The-Fly Native Ensemble[C]// NeurIPS 2018. 2018: 7517-7527.
[21] HOU S, PAN X, LOY C C, et al. Learning a Unified Classifier Incrementally via Rebalancing[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE Press, 2019: 831-839.
[22] ZHANG L, SONG J, GAO A, et al. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation[C]// Proceedings of IEEE International Conference on Computer Vision. Seoul: IEEE Press, 2019: 3712-3721.
[23] HOU Y, MA Z, LIU C, et al. Learning Lightweight Lane Detection CNNS by Self Attention Distillation[C]// Proceedings of IEEE International Conference on Computer Vision. Seoul: IEEE Press, 2019: 1013-1021.
[24] PARK S, KIM J, HEO Y S. Semantic Segmentation Using Pixel-Wise Adaptive Label Smoothing via Self-Knowledge Distillation for Limited Labeling Data[J]. Sensors, 2022, 22(7): 2623.
[25] KRIZHEVSKY A, HINTON G. Learning Multiple Layers of Features from Tiny Images[J/OL]. https://www.researchgate.net/publication/306218037_Learning_multiple_layers_of_features_from_tiny_images.
[26] NETZER Y, WANG T, COATES A, et al. Reading Digits in Natural Images with Unsupervised Feature Learning[J/OL]. https://www.researchgate.net/publication/266031774_Reading_Digits_in_Natural_Images_with_Unsupervised_Feature_Learning.
[27] XIAO H, RASUL K, VOLLGRAF R. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms[J]. arXiv: 1708. 07747, 2017.
[28] LE Y, YANG X. Tiny Imagenet Visual Recognition Challenge[S]. CS 231N, 2015.
[29] TUNG F, MORI G. Similarity-Preserving Knowledge Distillation[C]// Proceedings of IEEE International Conference on Computer Vision. Seoul: IEEE Press, 2019: 1365-1374.

WAN Xu, born in 1998, postgraduate, is a member of China Computer Federation. Her main research interest is knowledge graph.

MAO Yingchi, born in 1976, Ph.D, professor, Ph. D supervisor, is a senior member of China Computer Federation. Her main research interests include distributed data processing and edge intelligent computing.

(责任编辑:何杨)