计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2021/15(10)-1812-18
doi: 10.3778/j.issn.1673-9418.2104022

卷积神经网络压缩中的知识蒸馏技术综述

孟宪法,刘　方+,李　广,黄萌萌
国防科技大学 自动目标识别重点实验室,长沙 410000
+通信作者 E-mail: smartlf@sina.com

摘　要:近年来,卷积神经网络(CNN)凭借强大的特征提取和表达能力,在图像分析领域的诸多应用中取得了令人瞩目的成就。但是,CNN性能的不断提升几乎完全得益于网络模型的越来越深和越来越大,在这个情况下,部署完整的CNN往往需要巨大的内存开销和高性能的计算单元(如GPU)支撑,而在计算资源受限的嵌入式设备以及高实时要求的移动终端上,CNN的广泛应用存在局限性。因此,CNN迫切需要网络轻量化。目前解决以上难题的网络压缩和加速途径主要有知识蒸馏、网络剪枝、参数量化、低秩分解、轻量化网络设计等。首先介绍了卷积神经网络的基本结构和发展历程,简述和对比了五种典型的网络压缩基本方法;然后重点针对知识蒸馏方法进行了详细的梳理与总结,并在CIFAR数据集上对不同方法进行了实验对比;其后介绍了知识蒸馏方法目前的评价体系,给出多类型方法的对比分析和评价;最后对该技术未来的拓展研究给出了初步的思考。

关键词:卷积神经网络(CNN);知识蒸馏;神经网络压缩;轻量化网络
文献标志码:A　**中图分类号:**TP391.4

Review of Knowledge Distillation in Convolutional Neural Network Compression

MENG Xianfa, LIU Fang+, LI Guang, HUANG Mengmeng
National Key Laboratory of Science and Technology on Automatic Target Recognition, National Defense University of Science and Technology, Changsha 410000, China

Abstract: In recent years, convolutional neural network (CNN) has made remarkable achievements in many applications in the field of image analysis with its powerful ability of feature extraction and expression. However, the continuous improvement of CNN performance is almost entirely due to the deeper and larger network model. In this case, the deployment of a complete CNN often requires huge memory overhead and high-performance computing units (such as GPU) support. However, there are limitations in the wide application of CNN in embedded devices with limited computing resources and mobile terminals with high real-time requirements. Therefore, CNN urgently needs network lightweight. At present, the main ways to solve the above problems are knowledge distillation, network pruning, parameter quantization, low rank decomposition, lightweight network design, etc. This paper first introduces the basic structure and development process of convolutional neural network, and briefly describes and compares five typical basic methods of network compression. Then, the knowledge distillation methods are combed and summarized in detail, and the different methods are compared experimentally on the CIFAR data set. Furthermore, the current evaluation system of knowledge distillation methods is introduced. The comparative analysis and evaluation of many types of methods are given. Finally, the preliminary thinking on the future development of this technology is given.

收稿日期:2021-03-11　**修回日期:**2021-05-17

Key words: convolutional neural network (CNN); knowledge distillation; neural network compression; lightweight network

近些年来，随着数据集的丰富完善，计算单元性能的快速提升，卷积神经网络（convolutional neural network，CNN）凭借着强大的特征提取和表达能力，在图像分类[1]、目标检测[2]和语义分割[3]等计算机视觉领域，都获得了显著的应用成效。得益于网络模型的加深加巨，CNN的众多拓展模型在众多任务中超越了很多传统技术方法，甚至堪比人类的识别能力，可适用于广泛的行业应用，如智能驾驶[4-5]、智能机器人[6]、人脸识别[7]、疾病检测[8]等。

但是注意到：通常网络的性能与网络的结构复杂度成正比，网络结构越复杂模型越深越宽，网络性能就越好。如表1所示，He等提出的ResNet-152[9]模型深度达到了152层，参数量达到了0.6亿个，需要花费241 MB内存存储，分类一幅分辨率为224×224的彩色图像需要115亿次浮点型计算。深度神经网络会占用大量的内存存储，带来繁多的计算量，造成巨大的电量消耗，这些问题使网络很难在资源受限的嵌入式设备和对实时性要求较高的移动端部署。如果能把深度神经网络进行压缩，让网络减少对内存存储的消耗，就可以使网络在内存资源受限的设备上进行部署，并且轻量化的神经网络还可以实现运算加速，这样在军民应用中都有着广泛的拓展潜力。

Table 1 Basic information of classical convolutional neural network

表1 经典卷积神经网络的基本信息

网络名称	网络层数	参数量/10^6	内存大小/MB	Flops/10^9	MACC/10^9
AlexNet	8	61.10	244.4	0.72	1.43
VGGNet-16	16	138.36	553.5	15.50	30.96
VGGNet-19	19	143.67	574.7	19.67	39.28
GoogleNet	22	6.62	52.2	1.51	3.02
ResNet-50	50	25.58	102.5	4.12	8.22
ResNet-101	101	44.55	178.8	7.84	15.66
ResNet-152	152	60.19	241.6	11.57	23.11

通常情况下，深度神经网络存在着大量的参数冗余。根据LeCun等[10]实验表明，深度神经网络有近一半的权重对网络性能影响甚微，由此可见，深度神经网络有着很大的压缩空间，而且过量的参数还会导致过拟合，使模型的泛化能力下降。研究者们已经探索了网络剪枝（network pruning）[11-13]、参数量化（parameter quantification）[14-17]、低秩分解（low-rank decomposition）[18-19]、轻量化网络设计（compact structure design）[20-22]、知识蒸馏（knowledge distillation）[23-26]等方法。自AI教父Hinton提出了知识蒸馏技术后[23]，知识蒸馏受到了研究者们的广泛关注，并在网络压缩中展现了巨大的研究价值。目前网络压缩领域的综述性文章[27-31]缺乏专门对知识蒸馏技术的详细介绍，因此本文对卷积神经网络压缩中的知识蒸馏技术进行详细说明。

1 面向卷积神经网络的压缩方法简述

1.1 卷积神经网络的基本结构和发展历程

卷积神经网络是目前广泛应用的深度学习架构，它是一种层次模型，由用于特征提取的卷积层和用于特征处理的池化层交叉堆叠而成。卷积神经网络的输入为一般原始数据（如RGB图像、原始音频数据等），通过前馈运算来进行预测和推理，通过反馈运算来进行网络参数的训练和学习。

VGG[32]是一种经典的卷积神经网络，它采用卷积层和池化层交叉堆叠，最后连接全连接层的层次结构，其网络结构非常具有代表性。VGG一共有6个不同的版本，最常用的是VGG16，其网络结构如图1所示。VGG的核心思想是通过加深网络深度来提高网络性能，在2014年的ILSVRC挑战赛中表现优异，在定位比赛上取得了第一名的成绩，在分类比赛上以7.3%的top5错误率取得了第二名的成绩。

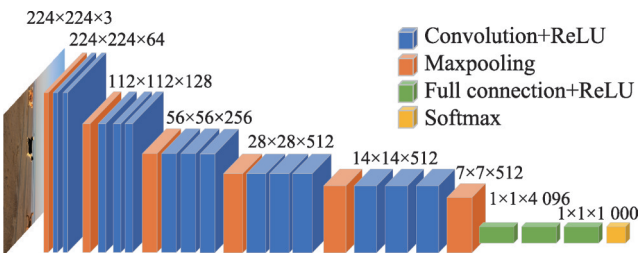

Fig.1 Stucture of VGG16 network

图1 VGG16网络结构

6个不同版本的VGG网络结构如表2所示。模型A-LRN在模型A的基础上多一个局部响应归一化层(local response normalization, LRN),但是实验表明,LRN对性能提升收效甚微,并且增加模型的内存占用和运行时间,因此后续模型都没有加入LRN。模型C在模型B的基础上增加了1×1卷积层,增加了网络的非线性表达能力。模型D用3×3卷积核代替了模型C的1×1卷积核,因为大感受野可以学习到更多的空间特征,增加了网络的学习能力。模型D对应VGG16,模型E对应VGG19。

Table 2 Structure of VGGNet network

表2 VGGNet网络结构组成

A	A-LRN	B	C	D	E
11层	11层	13层	16层	16层	19层
224×224 RGB image					
3×3×64	3×3×64 LRN	3×3×64 3×3×64	3×3×64 3×3×64	3×3×64 3×3×64	3×3×64 3×3×64
maxpooling					
3×3×128	3×3×128	3×3×128 3×3×128	3×3×128 3×3×128	3×3×128 3×3×128	3×3×128 3×3×128
maxpooling					
3×3×256 3×3×256	3×3×256 3×3×256	3×3×256 3×3×256	3×3×256 3×3×256 1×1×256	3×3×256 3×3×256	3×3×256 3×3×256 3×3×256
maxpooling					
3×3×512 3×3×512	3×3×512 3×3×512	3×3×512 3×3×512	3×3×512 3×3×512 1×1×512	3×3×512 3×3×512	3×3×512 3×3×512 3×3×512
maxpooling					
3×3×512 3×3×512	3×3×512 3×3×512	3×3×512 3×3×512	3×3×512 3×3×512 1×1×512	3×3×512 3×3×512	3×3×512 3×3×512 3×3×512
maxpooling					
FC-4 096					
FC-4 096					
FC-1 000					

卷积神经网络起源于20世纪60年代左右的神经科学领域中。LeCun在1998年提出基于梯度学习的改良版CNN模型LeNet-5[33]。在这之后,大量研究人员提出了很多方法去优化深层结构和克服深度神经网络在训练过程的困难,深度卷积神经网络的性能也因此得到了大幅提升。2012年卷积神经网络在ILSVRC2012挑战赛图像分类任务大放异彩,Krizhevsky提出的AlexNet[34]模型一举夺下2012年ILSVRC挑战赛冠军。继AlexNet之后,卷积神经网络迅速发展,陆续出现了很多性能优异的卷积神经网络模型,其中比较有代表性的有ZF-Net[35]、VGG[32]、GoogLeNet[36]、ResNet[9]、DenseNet[37]、DPN[38]、SENet[39]、MobileNetV1[40]、MobileNetV2[41]、SqueezeNet[42]和ShuffleNet[43]。

如图2所示,卷积神经网络的发展是由简到繁,不断发展的应用亟需网络在保持性能的前提下可高效适用于计算资源受限的平台上,探索可行的网络简化技术具有理论和应用双重意义。

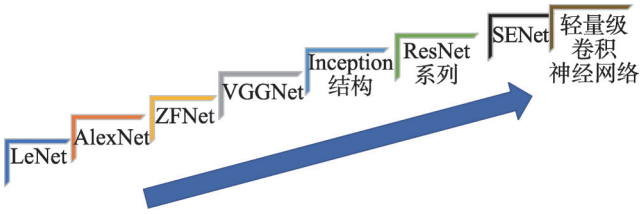

Fig.2 Development of convolutional neural network

图2 卷积神经网络的发展

1.2 知识蒸馏技术与其他压缩方法的对比

目前卷积神经网络压缩和加速的方法主要有以下五种:网络剪枝、参数量化、低秩分解、轻量化网络设计、知识蒸馏。为了综合示意以上方法特色,如表3所示,从方法的设计思想、作用位置、是否改动网络架构等方面进行对比,列举了以上方法的代表性研究工作,并对比分析了以上方法的优缺点。

知识蒸馏从充分发掘神经网络的性能潜力出发进行研究,旨在实现使用层级浅、结构简单的神经网络发挥良好的性能,相比于其他方法压缩出来的杂乱的结构,知识蒸馏可以选择任何一种网络结构整齐简洁的小型网络作为目标网络,并且不会改变小型网络的结构,知识蒸馏通过大型网络(即教师网络)辅助小型网络(即学生网络)训练的方式,提高学生网络性能,达到学生网络性能接近教师网络的效果,实现将教师网络压缩为学生网络的目的。通过知识蒸馏获得了层数更浅性能更好的小模型,小模型更适合部署因为它就是为了快而设计的,并且小模型并不需要花很多时间去调参,不需要特定的硬件就能直接实现模型加速。因此知识蒸馏逐渐发展为网络轻量化方法的一个热点分支。

Table 3 Comparison of network compression methods

表3 网络压缩方法对比

对比项目	网络剪枝	参数量化	低秩分解	轻量化网络设计	知识蒸馏
设计思想	设计参数评价标准以衡量参数重要性，移除不重要的参数	将网络中高精度的参数替换成低精度的参数	将原始张量分解为若干个低秩张量	采用紧凑高效的网络结构，设计适合部署在移动设备的网络	利用复杂度高的大型网络作为教师网络，用来指导复杂度较低的学生网络训练
作用位置	卷积层、全连接层	卷积层、全连接层	卷积层	整体网络	卷积层、全连接层
是否改动网络	需要	不需要	需要	需要	不需要
代表性研究	结构化剪枝、非结构化剪枝(group级别剪枝、filter级别剪枝)	二值化、三值化、聚类量化、混合位宽量化	二元分解、多元分解	卷积核级别(SqueezeNet、ShuffleNet、MobileNet系列)、层级别、网络结构级别(SplitNet、MorphNet)	输出层知识蒸馏、互信息知识蒸馏、注意力转移、相关性知识蒸馏、对抗性知识蒸馏
优点	非结构化剪枝可对网络进行任意程度的压缩；结构化剪枝可使网络变窄，便于在硬件上实现加速	可显著减少参数存储空间与内存占用空间，加快运算速度，降低设备能耗	在大卷积核和中小型网络上有不错的压缩和加速效果，方法研究已经比较成熟	网络训练简单，训练时间短，可获得存储量小、计算量低和网络性能好的小型网络，适宜部署到移动平台等资源受限设备	可以将大型网络压缩为任意小型网络，部署到移动平台等资源受限设备，易与其他压缩方法结合使用实现更大程度的压缩
缺点	非结构化剪枝造成网络结构不规整，难以有效加速；结构化剪枝可能会造成预测精度的下降	训练加微调耗时久；量化到特殊位宽时，易造成与硬件平台不兼容，灵活性差	难以分解精简卷积核和较小的卷积核，逐层分解不利于全局参数压缩	特殊结构很难与其他的压缩与加速方法组合使用；泛化性较差，不适合作为预训练模型帮助其他模型训练	至少需要训练两次网络，造成训练时间久

2 知识蒸馏方法

2014年，Hinton等人[23]首次提出了知识蒸馏(knowledge distillation，KD)的概念，并通过实验验证了其在卷积神经网络压缩上的有效性和可行性。知识蒸馏策略已在目标检测[44-46]、语义分割[47]、目标识别[48]、视频分类[49]、图像去雾[50]等很多计算机视觉应用中发挥了作用。

下文先给出基本分析原理，然后根据蒸馏位置的不同，把知识蒸馏划分为基于softmax输出层的知识蒸馏与基于中间层的知识蒸馏，其他代表性的方法还包括基于相关性知识的知识蒸馏以及结合生成对抗网络(generative adversarial networks，GAN)的知识蒸馏。

2.1 知识蒸馏的基本思想

知识蒸馏的本质体现在：老师会把自己的思考过程和总结作为一种知识精华传授给学生，学生通过理解学习，获得抽象、提炼后的知识，以达到和老师接近的水平。

知识蒸馏的基本思想正是让卷积神经网络模仿人类的学习行为，将大型网络(教师网)学习到的知识提炼传授给小型网络(学生网)，并指导小型网络的训练，从而实现了从大型网络压缩成小型网络的目的。其一般实现思路如图3所示。

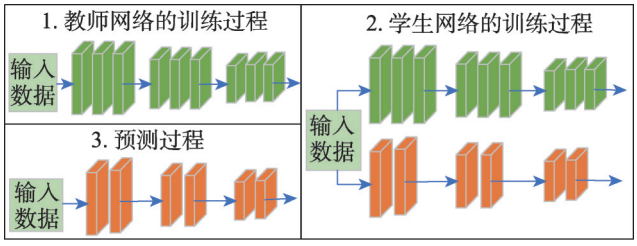

Fig.3 Realization of knowledge distillation

图3 知识蒸馏一般的实现思路

2.2 基于softmax输出层的知识蒸馏

基于softmax输出层的知识蒸馏(KD)[23]是由Hinton提出的知识蒸馏领域的开山之作。在一个训练成熟的网络模型输出的概率分布中，错误类别的概率一般比较小，但是其中可能会存在一些概率相对较高的类别，例如，一辆公交车可能只有很小的机会被误认为小汽车，但这个错误概率仍然比误认为一棵树的可能性高很多倍，Hinton等人认为这些错误类别的相对概率中隐藏着网络学习到的知识，这种知识是概率分布只有0和1的真实标签不具备的。

因此，Hinton提出KD，在softmax输出层中加入超参数T（如式(1)所示）用来平滑网络输出的概率分布，以强化输出的概率分布中网络学习到的知识。通过温度系数T平滑过后的网络输出被称为软目标，软目标和真实标签一起指导学生网络训练，损失函数的组成J_{KD}一般如式(2)所示。

$$\hat{q}_i=\frac{\exp(z_i/T)}{\sum_j\exp(z_j/T)} \tag{1}$$

$$J_{KD}=J_{CE}(y_{true},p)+\lambda T^2J_{CE}(\hat{p},\hat{q}) \tag{2}$$

神经网络通常通过使用softmax输出层来产生类别概率，该输出层通过将z_i归一化转化成概率q_i。$J_{CE}(y_{true},p)$表示学生网络的预测输出与真实标签的交叉熵，$J_{CE}(\hat{p},\hat{q})$表示学生网络平滑后的预测输出与教师网络平滑后的预测输出的交叉熵，λ为调节两个损失函数比例的超参数，因为经过超参数T平滑后的交叉熵，在反向传播时其梯度会变为原来的$\frac{1}{T^2}$，为了保持其梯度的尺度和真实标签对应的交叉熵的尺度一致，需要把平滑后的交叉熵乘以T^2。

下面介绍基于softmax输出层的知识蒸馏的拓展方法。

2.2.1 最高分差(TSD)知识蒸馏

研究动机：深度网络能够自动地为每幅图像学习语义相似的类，置信度越高的类在语义上更可能与输入图像相似。利用这些信息，可以让学生网络避免对不必要的严格分布进行拟合，从而获得更好的泛化能力。

方法实现：最高分差(top score difference，TSD)[51]在标签平滑正则化(label smoothing regularization，LSR)[52]和置信惩罚(confidence penalty，CP)[53]的基础上进行改进，TSD只使用教师网络预测输出置信度最高的k个类别计算损失，超参数k代表每个图像语义上最相似的类的数量，其中包含真实类别在内。然后，计算最高置信度类别与其之下得分最高的$k-1$类之间的置信度差距，将结果作为教师网络提供的损失，联合目标任务损失对学生网络进行训练。

2.2.2 提前停止知识蒸馏(ESKD)

研究动机：将神经网络部署在移动平台上一般需要较大的压缩率，学生网络和教师网络之间的规模差距也随之提升。虽然大模型的准确率更高，但是它往往并不能更好地指导学生网络训练，原因是容量不匹配。由于网络规模差距过大，学生无法模仿老师，反而会带偏了目标任务损失。这类问题的解决办法一般是采用分步蒸馏，从大模型提取到中模型，然后从中模型提取到小模型，但是这种分步的方法需要多次训练，造成训练时间也数倍地增长。

方法实现：文献[54]详尽地探索了影响知识蒸馏的因素，提出了另外一种思路——提前停止知识蒸馏(early-stopped knowledge distillation，ESKD)，在学生网络训练结束之前停止教师的知识指导以提高学生网络学习效果，通过使教师网络知识对应的损失权重逐渐衰减，以获得一个更适合学生训练的方案。

2.3 基于中间层的知识蒸馏

基于中间层的知识蒸馏是目前研究最多的方法[55-60]，主要是从网络的中间隐藏层中通过各种手段提取可以表示网络学习过程的知识，或者能够蕴藏网络如何对输入数据进行推理的知识，将提取到的这些知识传递给学生网络以实现知识蒸馏，达到提高学生网络性能的目的。下面介绍一些典型的方法。

2.3.1 FitNet

研究动机：增加网络深度可以重复使用特征，获得在更高层次上更抽象和不变的特征表示。受此启发，FitNet[55]使用更深更窄的学生网络和更浅更宽的教师网络以实现更好的蒸馏效果，并且同时使用教师网络的软目标和教师网络的中间层特征图作为知识。

方法实现：使用教师网络的特征图作为指导层，选择学生网络的特征图作为被指导层，FitNet是一个二阶知识蒸馏，第一步使用指导层指导被指导层训练，损失函数如式(3)所示，第二步使用KD[23]继续训练学生网络。

$$L_{HT}(W_{Guided},W_r)=\frac{1}{2}\|u_h(X;W_{Hint})-r(v_g(X;W_{Guided});W_r)\|^2 \tag{3}$$

其中，u_h、v_g和r分别代表教师网络、学生网络和适配器的嵌套函数，X是输入特征图，W_{Hint}、W_{Guided}和W_r分别代表教师网络的权重、学生网络的权重和适配器的权重。

直接由卷积层生成的特征图通常尺寸较大，计算成本高，而且学生网络很难学习。为了解决这个问题，Lee等人[56]提出结合奇异值分解(singular value decomposition，SVD)的知识蒸馏，通过减少特征地图的空间维数，有效去除特征映射中的空间冗余，在特征降维过程中获得有意义的隐含特征信息，并将这

种信息传递给学生网络。

2.3.2 注意力转移(AT)

研究动机：FitNet要求学生模拟教师的全部特征图，这样的要求太严格。文献[24]提出了注意力转移(attention transfer，AT)来放宽FitNet的假设，注意力图是对多个通道的特征图的总结，使用一个注意力图来代替多通道的特征图。

注意力机制：注意力是视觉体验的一个关键因素，并与感知密切相关，注意力的集中程度体现了重视程度，人类需要保持注意力，以建立一个具有细节和连贯性的视觉表现。受此启发，人工注意力的核心思想是，通过集中注意力让系统更关注一个对象或区域，以更详细地检查它。

方法实现：AT把注意力看作一组空间映射，这些映射可以在网络的各个层中定义，以便它们能够捕获低、中、高级的表示信息，然后把注意力从教师网络转移到学生网络，以提高后者的表现。如图4所示，AT定义了基于激活值(神经元在预测过程的输出)的注意力图，其基本假设是隐藏层神经元激活的绝对值可以代表这个神经元的重要性，通过对同一空间位置不同通道的特征图的统计，将 C 个通道的特征图映射为单通道的注意力图。通过让学生网络的注意力图拟合教师网络的注意力图，并联合目标任务损失对学生网络进行训练。

文献[25]提出通过匹配注意力图和它们的雅可比矩阵进行知识蒸馏，该蒸馏方法的教师和学生网络结构可以是任意的。该方法主要利用了神经网络雅可比矩阵两个重要的性质：第一，维度与网络结构无关，只与输入和输出的维度有关，因此，不同网络的雅可比矩阵可以进行比较；第二，对于相同的网络，不同的权重配置可能得到相同的雅可比矩阵，这是由于网络的冗余性和损失函数非凸性造成的。

2.3.3 FSP蒸馏

研究动机：卷积神经网络是一个层次结构，特征从输入到输出逐层传递，神经网络学习到的知识可以定义为如何构建一个从输入到输出的映射关系，进一步可以分解为层与层之间的特征变换关系，如图5所示。FSP(flow of solution procedure)蒸馏[57]将这种层与层之间的特征关系从教师网络传递给学生网络。单纯地让学生网络模仿老师网络生成的特征图是硬约束，会让学生网络变得不灵活，因此更好的办法是教会学生学习的过程，FSP蒸馏定义了FSP矩阵来刻画层与层之间的特征关系，其核心思想便是授人以鱼不如授人以渔。

方法实现：对于具有相同尺寸的特征图，使用低层和高层不同通道的特征图两两计算内积，得到的结果代表对应通道的两两特征图的互相关值，将互相关值作为FSP矩阵的对应位置的元素。FSP矩阵的计算过程如式(4)所示。最后用 L_2 损失去拉近教师和学生的FPS矩阵之间的距离，通过构建如式(5)所示的FSP损失联合目标任务损失一起指导学生网络训练。FSP蒸馏的概念图如图5[29]所示。

$$G_{i,j}(x;W)=\sum_{s=1}^{h}\sum_{t=1}^{w}\frac{F^1_{s,t,i}(x;W)\times F^2_{s,t,i}(x;W)}{h\times w} \tag{4}$$

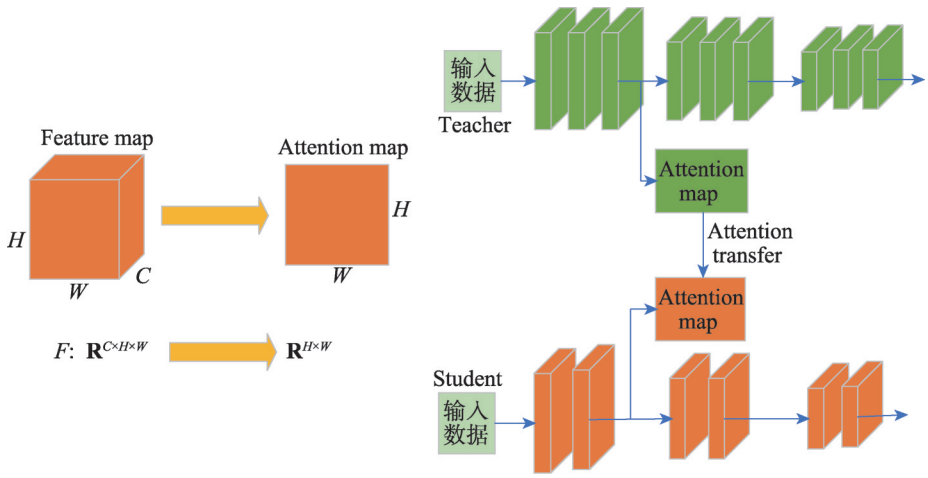

Fig.4 Concept map of knowledge distillation for attention transfer

图4 注意力转移知识蒸馏概念图

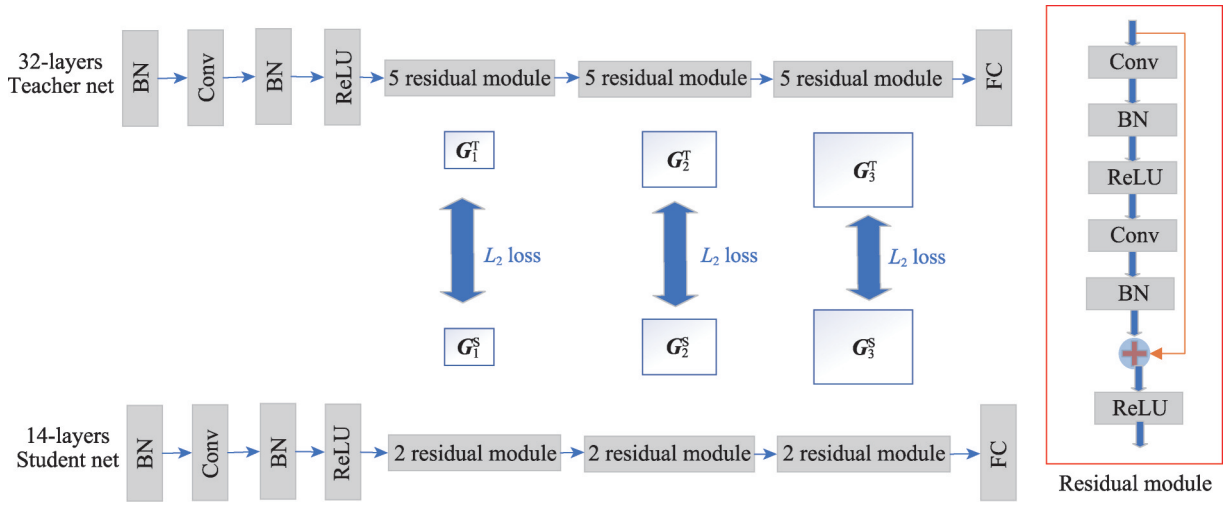

Fig.5 Concept diagram of FSP distillation

图5 FSP蒸馏的概念图

$$L_{\mathrm{FSP}}(W_t,W_s)=\frac{1}{N}\sum_{x}\sum_{i=1}^{n}\lambda_i\times\|(G_i^{\mathrm{T}}(x;W_t)-G_i^{\mathrm{S}}(x;W_s))\|_2^2 \quad (5)$$

其中，F_1 和 F_2 分别表示低层和高层特征图；h 和 w 分别表示特征图的长和宽；i 和 j 分别表示低层和高层特征图的通道索引；x 和 W 分别表示输入和参数。

2.3.4 神经元选择性转移(NST)

研究动机：文献[58]提出在知识蒸馏过程中，直接对特征图进行匹配并不是最好的选择，因为它忽略了样本密度，并提出采用分布对齐的神经元选择性转移(neuron selectivity transfer，NST)方法。

方法实现：NST的假设是每个神经元都从原始输入中提取与目标任务相关的特定模式，因此，如果一个神经元在某些区域或样本中被激活，这就意味着这些区域或样本可能具有一些与任务相关的共同属性。NST通过匹配教师和学生网络之间神经元选择性模式的分布进行知识蒸馏，采用最大平均偏差(maximum mean discrepancy，MMD)作为损失函数来度量师生分布之间的差异，并结合目标任务损失对学生网络进行训练。

2.3.5 互信息知识蒸馏

研究动机：基于互信息的知识蒸馏[59,61]使用的手段是最大化教师和学生网络之间的互信息，学生网络通过学习教师网络中激活值的分布最大化互信息，从而进行知识的传递。在学生网络已知的条件下，当教师网络的熵很小时，这说明学生网络已经获得了能够拟合教师网络所需要的知识，因此学生网络的性能也已经接近教师网络。如式(6)所示，在 $H(t)$ 已知的条件下，$H(t/s)$ 的值越小时，互信息 $I(t;s)$ 越大。

$$I(t;s)=H(t)-H(t/s)=-E_t[\lg p(t)]+E_{t,s}[\lg p(t/s)] \quad (6)$$

由于互信息的计算较困难，变分信息蒸馏(variational information distillation，VID)[31]采用变分信息最大化方案来最大化变分下界，如式(7)所示，即用一个可变高斯分布 $q(t/s)$ 来模拟 $p(t/s)$，由于蒸馏过程中 $H(t)$ 和需要学习的学生网络参数无关，最大化互信息就转换为最大化可变高斯分布的问题。

$$\begin{aligned}I(t;s)&=H(t)-H(t/s)=H(t)+E_{t,s}[\lg p(t/s)]=\\&H(t)+E_{t,s}[\lg q(t/s)]+E_s[D_{\mathrm{KL}}p(t/s)\|q(t/s)]\geqslant\\&H(t)+E_{t,s}[\lg q(t/s)]\end{aligned} \quad (7)$$

方法实现：学生网络通过最小化与真实标签的交叉熵损失，同时与教师网络保持高度的互信息以学习教师网络的知识。

2.3.6 因子传输(FT)

研究动机：当教师网络和学生网络在网络结构、通道数量和初始条件等差距较大时，学生网络不能很好地理解教师网络特征图中复杂的知识，受此启发，文献[60]提出使用通道扩展的方法——因子传输(factor transfer，FT)蒸馏去进一步解释教师的知识，以帮助学生网络学习。

方法实现：先将教师网络特征图的通道扩展 k 倍，使 m 个通道特征图的知识转化到 $m\times k$ 个通道上，再进行知识传递。整体架构如图6[60]所示，在教师网络指导层的特征图后面连接一个额外的释义器模

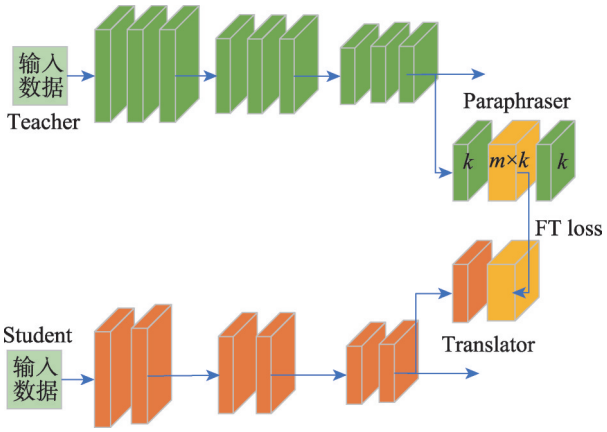

Fig.6 Concept diagram of FT distillation

图6 FT蒸馏的概念图

块，将特征图通道数扩展为 $m\times k$。释义器由卷积模块组成，为了保证通道扩展后得到的信息可以完整地表示原始特征图，在释义器的输入特征 x 和输出特征 $p(x)$ 间设置重建损失，对释义者进行无监督训练。相应地，在学生网络的被指导层后连接一个适配器，适配器也由卷积模块构成，目的是为了让学生网络的通道数与释义器的输出相匹配。通过式(8)计算蒸馏损失，其中，F_T 和 F_S 分别表示释义特征和适配器特征。

$$L_{FT}=\left\|\frac{F_T}{\|F_T\|_2}-\frac{F_S}{\|F_S\|_2}\right\|_P \tag{8}$$

2.3.7 最佳指导路径

研究动机：对于基于中间层的知识蒸馏，一个重要问题是如何确定最佳指导路径，即确定教师网络中的哪一层作为指导层和学生网络中的哪一层作为被指导层，怎么能使学生网络获得最好的指导效果。如图7所示，文献[62]提出一个迭代剪枝的优化方案来寻找最佳指导路线。

方法实现：将所有教师网络和学生网络特征图尺寸一致的层构成的指导路径作为指导路径集合，在训练超参数相同的情况下遍历所有可能的路径，动态确定最佳指导路径。对学生网络的特征图使用 1×1 的卷积核进行分析，这个操作可以降低特征图的维度和计算复杂度，并过滤出独特的通道特征。经过 1×1 的卷积核提取的新特征图用来与教师网络的特征图计算损失，损失函数如下所示：

$$L_{TP}(W_S,W_T)=h[u(x;W_T);W_{TP}]\times\|r[u(x;W_T);W_{TP}]-v[x;W_S]\| \tag{9}$$

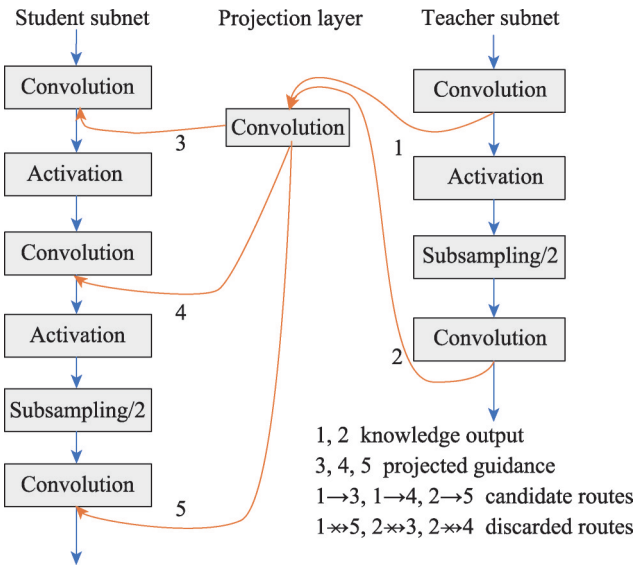

Fig.7 Candidate paths of knowledge transfer

图7 知识传递的候选路径

$$h(x)=\begin{cases}1, & x\geqslant 0\\ \eta, & x<0\end{cases} \tag{10}$$

其中，u 和 v 分别表示指导层和被指导层的深度嵌套函数，W_T 和 W_S 分别表示教师网络参数和学生网络参数。r 是应用在 u 上的知识投影函数，参数 W_{TP} 是卷积适配层的参数。u、v 和 r 必须在空间维度上具有可比性。

2.4 其他方法

2.4.1 相关性知识蒸馏

(1)关系知识蒸馏

研究动机：文献[63]提出了个体知识蒸馏(individual knowledge distillation，IKD)和关系知识蒸馏(relational knowledge distillation，RKD)的概念，IKD[23-24,51,55,60,64]使用单个输入样本在网络特征提取过程中生成的特征图或网络的输出作为知识进行蒸馏，使学生网络的输出模拟教师网络的输出，以此来模拟大模型的拟合能力。IKD中每个输入样本都是独立的，学生网络只能学习教师网络对单个输入样本的推理过程和输出结果，无法学习到多个输入样本在教师网络特征空间的相关性，这种相关性包含了教师网络对类内样本的聚合能力和类间样本的区分能力，以及教师网络的结构信息。

方法实现：如图8[63]所示，RKD算法的核心是以教师网络的输出为结构单元，取代IKD中以教师网络单个输出为知识的蒸馏方式，RKD利用多输出组

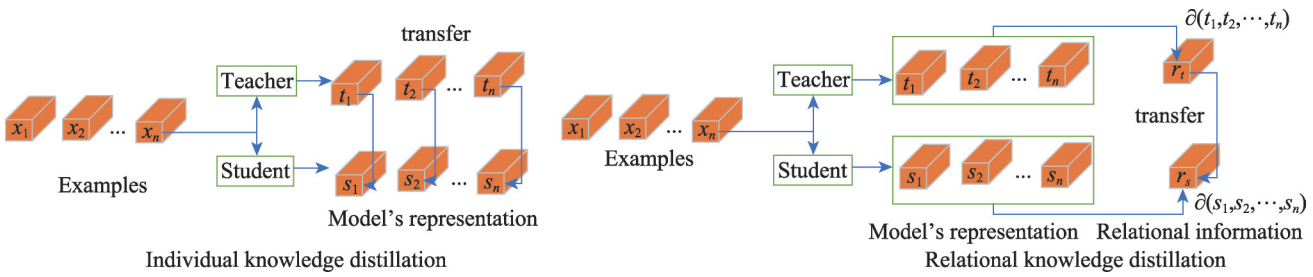

Fig.8 Individual knowledge distillation and relational knowledge distillation

图8 个体知识蒸馏与关系知识蒸馏

合成结构单元，更好地获取教师网络的结构化特征。RKD学习的损失函数如式(11)所示，其中 $t_1, t_2, \cdots, t_n$ 表示教师网络的多个输出，$s_1, s_2, \cdots, s_n$ 表示学生网络的多个输出，∂ 是构建结构信息的函数，由两个样本之间的欧几里德距离或三元组之间的角距离实现，l 表示计算二者之间的差距。

$$L_{\mathrm{RKD}}=\sum_{(x_1,x_2,\cdots,x_n)\in X^N} l(\partial(t_1,t_2,\cdots,t_n),\partial(s_1,s_2,\cdots,s_n)) \quad (11)$$

(2)样本关系图蒸馏

研究动机：知识蒸馏的主要挑战是如何从教师网络中提取一般的、适度的、充足的知识来指导学生网络。文献[65]提出了一种用于知识提取的样本关系图(instance relationship graph，IRG)，它对样本特征、样本关系和特征空间变换这三种知识进行建模，其概念图如图9[58]所示。

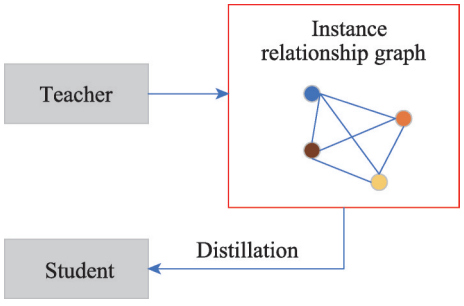

Fig.9 Concept diagram of instance relationship graph distillation

图9 样本关系图蒸馏示意图

方法实现：首先，构造IRG，其中顶点表示训练样本，两个顶点之间的边权重表示样本之间的相似性程度。然后，使用IRG变换对从低层到更高层的特征空间转换进行建模，尽管输入样本的特征在不同的网络结构中通常具有不同的维度，但图的大小总是相同的，因为节点的数量等于单次输入训练样本的数量。最后，设计IRG损失、IRG变换损失以及目标任务损失的联合损失函数，使用联合损失函数指导学生网络训练。

(3)图知识蒸馏

使用图作为网络中间层的拓扑表示，可以用来解释网络正在学习什么[66-67]或增强其鲁棒性[68]。图知识蒸馏(graph knowledge distillation，GKD)[69]可以看作是RKD欧几里德版本的拓展。在RKD的基础上，利用图捕捉隐藏空间的几何特征，建模多个样本之间的关系距离，提取相关性知识，进行知识蒸馏。GKD使用余弦距离来度量样本相似性，余弦距离相对于欧氏距离，更多的是从方向上区分差异，而对绝对的数值不敏感，可以更准确地衡量样本特征之间的相似性。为了避免过分重视离群值，对邻接矩阵进行了规范化处理，使用huber损失作为损失函数。

文献[70]在此基础上对损失函数进行了扩展，提出了三元组蒸馏(triplet distillation)，使用三元组损失[71]，通过自适应地改变正负对之间的距离，将相似信息从教师网络转移到学生网络。

(4)相似性保留知识蒸馏(SPKD)

研究动机：语义相似的输入往往会在一个经过训练的神经网络中产生相似的激活模式，文献[65]提出了相似性保留知识蒸馏(similarity-preserving knowledge distillation，SPKD)，并提出输入样本之间的相似性反映了教师网络在特征空间中表示特征的规律，有助于在特征空间中减小类内间距和增大类间间距，以提高学生网络的学习效果。SPKD使学生网络不用去模仿教师网络提取到的特征，只需要在自己的特征空间中保持样本之间的相似性与教师一致即可。其概念图如图10[72]所示。

方法实现：对于输入的样本数为 b 个的小批量图像，选择某一层特征图计算形状为 $b\times b$ 的相关矩阵，

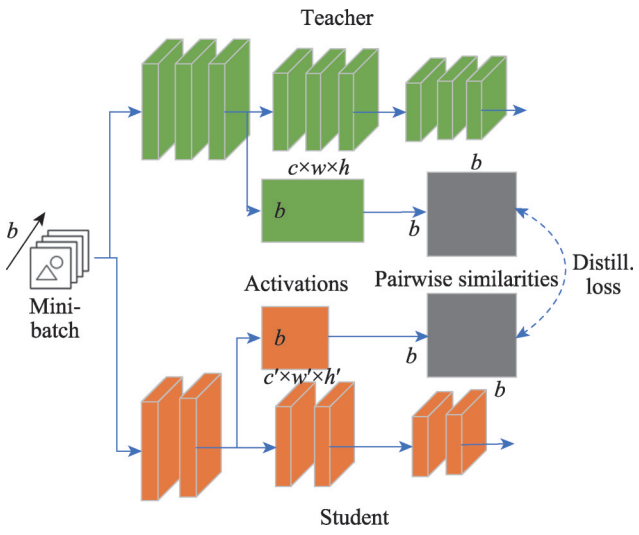

Fig.10 Similarity preserving knowledge distillation

图10 相似性保留知识蒸馏

蒸馏损失定义在学生和教师产生的相关矩阵上。教师和学生网络生成的相关矩阵的形状都是 $b\times b$，因此不用刻意保证师生网络的特征图大小和通道数相等，放宽了教师网络和学生网络的选择范围，方法实现了更好的泛化。

(5)相关同余知识蒸馏(CCKD)

文献[73]提出了相关同余知识蒸馏(correlation congruence for knowledge distillation，CCKD)，将教师网络对输入样本的预测输出和样本间相似性知识都传递给学生网络。此外，为了更好地捕捉样本之间的相关性，CCKD使用高斯径向基函数(Gaussian-RBF)去衡量样本之间的相关性，该函数如下所示：

$$k(x,y)=\exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right) \tag{12}$$

高斯径向基函数是一种常用的核函数，其值只依赖于距离原点空间的欧氏距离。相比于双线性池化[74]，高斯径向基函数在捕捉样本之间复杂的非线性关系方面更为灵活和强大。

CCKD的采样策略：样本之间的相关性是在小批量输入中计算的，因此一个合适的采样器对于平衡类内的一致性和类间的相关性非常重要。一种简单的策略是均匀随机抽样，但是当类数较多时，所有样本都属于不同的类，这会导致类内相关性梯度的高偏差估计。为了解决这个问题，CCKD采用两种小批量采样器策略：类均匀随机采样器和超类单形随机采样器。均匀随机抽样按类抽样并随机为每个抽样类别选择固定 k 个样本数(例如，一个批次中包含5个类别，每个类别包含 $k=8$ 个样本，组成一个包含40个样本的批次)。超类单形随机采样器与类均匀随机采样器相似，不同之处在于，它通过超类对样本进行采样，超类是通过聚类生成的真实类的一种更软的形式。为了得到训练样本的超类，首先使用教师网络提取特征，然后使用 K-均值聚类，样本的超类被定义为它所属的集群。由于超类改变了特征空间中样本的粗糙结构，超类单形随机采样器比类均匀随机采样器更灵活，更能容忍不平衡标记。

(6)多头图蒸馏(MHGD)

文献[75]提出了多头图蒸馏(multi-head graph distillation，MHGD)方法获取输入样本的嵌入知识。首先提取网络两个层对应的特征映射，使用KD-SVD[56](knowledge distillation using singular value decomposition)通过径向基函数将特征映射压缩为特征向量，然后将小批量样本输入教师网络，生成两个特征向量集。通过计算两个特征向量集之间的关系提取知识，使用传递知识的损失联合目标任务损失同时指导学生网络训练。

2.4.2 基于生成对抗网络的知识蒸馏

研究动机：基于生成对抗网络(GAN)[26]的知识蒸馏，利用生成对抗策略以实现知识从教师网络到学生的传递。当学生网络比教师网络小很多时，强迫学生网络精确模拟教师网络是很困难的，而GAN有助于保持输出分布的多模态性质[60]，并减小手工调参的误差，学生网络可以自动学习到良好的损失，转移类间的相关性，提升学生网络的性能。

方法实现：对抗知识蒸馏[76-78]将学生网络作为生成器，先分别获取生成器和教师网络生成对输入样本的输出概率分布，再使用判别器来区分学生网络的输出与教师网络的输出。学生网络和判别器交替更新参数，其中判别器的更新为了更好地区分教师网络的输出与学生网络的输出，学生网络的更新为了更好地欺骗判别器，使判别器无法区分学生网络的输出与教师网络的输出。经过学生网络和判别器的多次交替更新后，并达到使判别器无法区分学生网络的输出与教师网络的输出的效果，最终实现学生网络模拟教师网络的目的。图11描述了基于生成对抗网络的知识蒸馏的一般思想。

虽然判别器捕获了教师和学生输出的高级统计数据，但是缺少低级对齐，并且对抗性训练过程很困

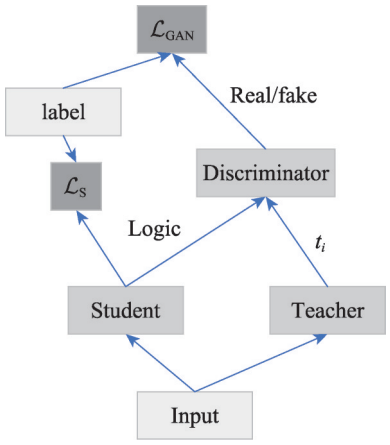

Fig.11 Knowledge distillation based on generative adversarial network

图11 基于生成对抗网络的知识蒸馏

难[76]。文献[79]提出了更严格的训练方式，其增加了判别器的预测目标，判别器除了预测真、假之外，还预测类别标签，使对抗训练变得更加稳定，鉴别器还可以在学生和教师的输出之间提供类别级的对齐，判别器的输出是一个 $C+2$ 维向量，带有 C 个标签预测和一个真/假预测。文献[77]提出了KDGAN(knowledge distillation with generative adversarial networks)的框架，利用生成对抗网络提取知识用于多标签学习。KDGAN框架定义为一个极大极小博弈，其中学生网络、教师网路和判别器对立地训练，极大极小博弈具有均衡性，有利于学生网络更好地模拟真实数据的分布。在对抗性训练中，利用具体的分布来控制梯度的方差，得到低方差的梯度估计以加速训练。

2.5 知识蒸馏目前的评价体系

目前知识蒸馏的评价体系尚未完善，本文在展望部分针对知识蒸馏评价标准的规范化提出一些建议，下面仅介绍目前知识蒸馏普遍使用的评价方法。

目前知识蒸馏的评价是通过图像分类任务实现的，使用的数据集包括CIFAR-10、CIFAR-100、ImageNet、SVHN、ILSVRC2012、MNIST。其中CIFAR-10、CIFAR-100为最常用的数据集。

CIFAR-10是一个普适物体的小型彩色图像数据集。一共包含10个类别的RGB彩色图片：飞机、汽车、鸟类、猫、鹿、狗、蛙类、马、船和卡车。每个图片的尺寸为32×32，每个类别有6 000个图像，数据集中一共有50 000张训练图片和10 000张测试图片。

CIFAR-100包含100个类别的RGB彩色图片，每个类包含600个图像，每类各有500个训练图像和100个测试图像。CIFAR-100中的100个类被分成20个超类。每个图像都带有一个“精细”标签(它所属的类)和一个“粗糙”标签(它所属的超类)。

评价指标包括3个：(1)学生网络相对于教师网络参数的减少量；(2)通过知识蒸馏后的学生网络相比于教师网络准确率的降低量；(3)通过知识蒸馏后的学生网络相比于正常训练的学生网络准确率的提升量。第一个指标体现了算法对网络的压缩程度；第二个指标体现了在第一个指标对应的压缩程度下，算法对准确率的损失程度；第三个指标体现了在第一个指标对应的压缩程度下，该知识蒸馏算法的有效程度。

评价流程：(1)在实验数据集上训练多个教师网络(如ResNet152、ResNet101、WRN40-2等)，训练多个教师网络的目的是为了后续测试算法对网络结构和复杂度的泛化性。(2)对每一个教师网络，都选择多个不同复杂度的学生网络(如ResNet50、WRN40-1、WRN16-2、WRN4016-1等)进行蒸馏训练，学生网络的训练与教师网络使用相同的数据集，以进行准确率的对比，选择多个不同复杂度的学生网络的目的是为了测试算法对压缩程度的鲁棒性。一般情况下，当压缩程度超出一定范围时，蒸馏的效果会急剧下降。(3)使用正常的训练方式，对学生网络在相同的数据集上进行训练，以测试蒸馏的有效性。

2.6 知识蒸馏技术的对比实验与评价

2.6.1 不同知识蒸馏技术的对比实验

表4展示了各种知识蒸馏方法的性能，实验在CIFAR-100数据集上对WRN40-2进行两种程度的压缩，将WRN40-2压缩为WRN16-2和WRN40-1。在两种程度的压缩中，KD表现均最优；在WRN40-2到WRN16-2的压缩中，AT表现第二好；在WRN40-2到WRN40-1的压缩中，VID表现第二好。

表5展示了不同知识蒸馏方法与KD组合表现的性能，实验在CIFAR-100数据集上进行。将WRN40-2压缩为WRN16-2，有6种知识蒸馏方法与KD组合表现的性能超越KD单独使用的性能，其中AT+KD表现最优，这表明这6种方法提取到了KD缺乏的知识，与KD存在互补关系。

2.6.2 对各种知识蒸馏技术的分析评价

知识蒸馏的发展关系如图12所示，其中第一部

Table 4 Experiment of different knowledge distillation methods on CIFAR-100

表4 不同知识蒸馏方法在CIFAR-100上的实验

Teacher Net		WRN40-2	WRN40-2
Student Net		WRN16-2	WRN40-1
Teacherparms		2 248 954	2 248 954
Studentparms		693 498	566 650
Teacheracc/%		75.62	75.62
Studentacc/%		73.21	71.92
Accuracy/%	KD	74.91	73.54
	FitNet	73.57	72.24
	AT	74.09	72.75
	FSP	72.91	不适用
	NST	73.62	72.22
	VID	74.07	73.33
	FT	73.22	71.58
	RKD	73.35	72.19
	SPKD	73.84	72.44
	CCKD	73.59	72.23

Table 5 Combination experiment of different knowledge distillation methods and KD

表5 不同知识蒸馏方法与KD组合实验

Method	Accuracy/%	Method	Accuracy/%
KD	74.91	KD	74.91
FitNet	73.57	FitNet+KD	75.19
AT	74.09	AT+KD	75.33
NST	73.62	NST+KD	74.67
VID	74.07	VID+KD	75.12
FT	73.22	FT+KD	75.15
RKD	73.35	RKD+KD	74.89
SPKD	73.84	SPKD+KD	75.01
CCKD	73.59	CCKD+KD	75.11

分为基于softmax输出层的知识蒸馏方法，第二部分为基于中间层的知识蒸馏方法，第三部分为相关性知识蒸馏，第四部分为对抗性知识蒸馏。下面对上述四种类型的知识蒸馏方法的优缺点进行分析评价，并对其拓展研究提供一些思路：

（1）基于softmax输出层的KD

优点：KD能够有效地将大型教师网络压缩为小型学生网络，实现思路简单，适用于任意网络结构，在多分类任务中表现了优秀的性能。

缺点：该方法也存在着一定的局限性，由于它是基于softmax层输出的概率分布，严重依赖类的数量，

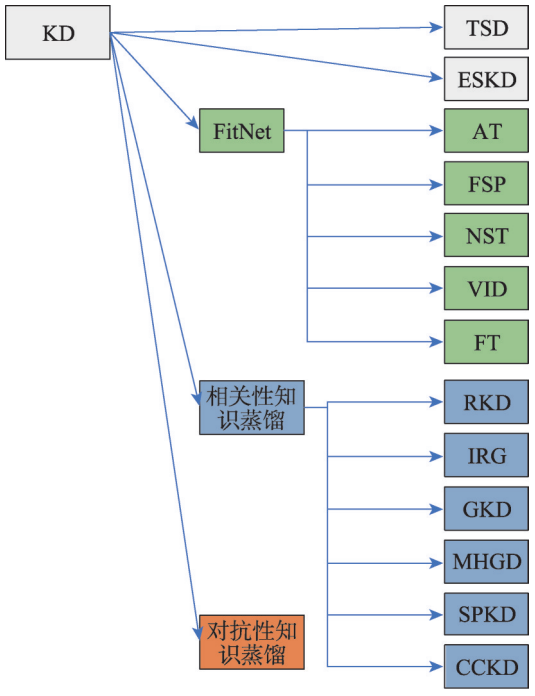

Fig.12 Development of knowledge distillation

图12 知识蒸馏的发展关系

因此应用场景只能局限于多分类问题，在二分类甚至目标类别较少的情况以及回归问题的表现并不理想；并且softmax输出层位于网络的最后，特征经过多次池化降维，所包含的信息量较少，提取到的知识语义程度太高，单纯地拟合教师网络的输出可能会造成学生网络的过拟合，影响学生网络的学习效果。

拓展研究：KD在知识蒸馏的快速训练方向有着良好的发展潜力，由于KD传递的知识位于网络的输出层，KD可以通过多模型集成的思路，将多个小模型同时训练，把小模型的预测输出通过集成产生知识，同时使用知识对其中的小模型进行指导，从而实现一阶段训练的知识蒸馏，达到快速训练的效果；探索可以与KD结合使用的中间层知识蒸馏方法，实现对输出层知识的补充，从而达到更好的蒸馏性能。

（2）基于中间层的知识蒸馏

优点：大量的研究人员提出了各式各样的知识提取方法，包括AT、FSP、NST、VID、FT等，基于中间层的知识蒸馏取得了较好的效果，相对于基于softmax输出层的知识蒸馏，其拥有更丰富多样的知识提取和传递的手段，同时也拥有更广阔的发展潜力和研究价值，其应用范围也取得了极大的扩展，不再依赖分类任务以及类的数量，可以应用在检测和分割

任务上。通过实验发现,AT实现了最好的性能,FSP和FT可以帮助学生网络提高收敛速度。

缺点:由于中间层的特征维度比输出层庞大很多,因此加剧了训练难度;并且在学生网络固定的时候,很难确定最优的指导路线以及最佳的指导教师,增加了人工调参的难度,可能需要多次调参才能获得训练最佳的学生网络;基于中间层特征图的知识蒸馏方法实现相对复杂,AT需要教师网络和学生网络的特征图适配,实际应用中可能会需要额外的适配操作,FT需要提前对释义器进行额外的训练,同时释义器和适配器会增加额外的参数;基于中间层知识蒸馏的研究动机大多从试探的角度出发,该方向的研究缺乏明确的理论知识指导,需要较高创新直觉。

拓展研究:基于中间层的知识蒸馏是目前知识蒸馏领域的热点研究,该方向可以结合神经网络最新的研究成果,探索更有效的知识提取和传递的手段。在AT的基础上可以探索结合更多类型的注意力机制的蒸馏方法,如空间注意力机制、通道注意力机制、混合型注意力机制以及自注意力机制等;如何选择教师到学生的指导路径也是重要的研究方向。

(3)相关性知识蒸馏

优点:相关性知识蒸馏有利于减少特征空间中的类内间距,扩大类间间距,对不同的网络结构具有较强的鲁棒性。

缺点:相关性知识蒸馏传递的知识太杂乱,各个损失分量之间的权重比例缺乏规律性,需要多次调参和试错才能找到较好的训练效果;由于相关性知识需要从批量数据中提取,输入数据的类别分布对蒸馏的效果也有较大的影响。

拓展研究:相关性知识蒸馏从样本的角度出发,因此,对输入样本进行预处理和采样策略可以作为重点研究方向。

(4)基于生成对抗网络的知识蒸馏

优点:基于生成对抗网络的知识蒸馏提供了知识蒸馏的一种全新的思路,通过使用鉴别器和学生网络交叉迭代训练使学生网络不断向教师网络靠近,该方法可以实现端到端的训练,解决了人工调参带来的误差和麻烦,而且训练效果较好。

缺点:由于这种训练方式使学生网络仅仅模拟教师网络的输出端,无法学习到教师网络内部更丰富的知识,并且鉴别器和学生网络的交叉迭代训练造成网络收敛速度慢,训练时间长。

拓展研究:基于生成对抗网络的知识蒸馏作为知识蒸馏的一个新的领域,以对抗训练作为训练手段,探索可以结合网络中间层知识的方法,以实现更好的蒸馏效果和更快速的训练。

3 展望知识蒸馏的拓展方向

随着深度神经网络研究的不断深入以及对神经网络网络轻量化日益扩大的应用需求,知识蒸馏技术在未来一定会大放异彩,未来的研究工作可参考如下几方面。

(1)更有效的知识提取

在知识蒸馏中,最重要的一个环节就是如何从教师网络中提取能够有效地指导学生网络进步的知识,基于注意力机制和相关性等提取知识的方法展现了不错的效果,但提取到的知识还是不够准确和有效,学生网络需要更精准的教师网络知识以获得更好的指导效果,因此如何改进知识提取的方法有很大的研究意义也是目前研究亟需解决的问题。

(2)知识蒸馏技术的研究领域扩展

现阶段知识蒸馏技术的研究主要是在分类问题上的,对于检测、分割等复杂任务的应用上还存在一定的局限性。对于同样的知识蒸馏算法,应用在分类问题上的压缩有着很好的效果,而应用到目标检测任务时性能可能会大幅降低,主要有以下几点原因:基于输出层的知识蒸馏方法是针对分类问题提出来的,它的假设是所有的输出类别有着相同的重要性,但是目标检测任务中,通常背景类所占比例要远高于目标类;检测任务相对于分类任务更复杂,需要同时处理目标分类和边框回归问题,对网络能力有更强的要求;检测任务主要关注物体真实标签重叠的局部区域,而分类模型更关注全局背景,在检测任务中,知识蒸馏将整个教师网络的知识提取给学生网络,其中包含了学生不需要的知识,冗余的知识会影响蒸馏的效果。因此知识蒸馏技术在目标检测的应用还存在着许多问题需要克服,这也正是知识蒸馏技术未来重要的研究方向。

(3)与网络剪枝技术等策略结合

基于知识蒸馏的网络压缩技术使小型网络超越自己正常的训练极限,达到与大型网络相当的准确率,但是经过知识蒸馏训练完成的小型网络的参数

仍然存在大量冗余，网络剪枝技术刚好可以很好地解决这个问题，通过知识蒸馏技术和网络剪枝技术的联合使用可以更大程度地压缩网络，因此如何将两种压缩技术不冲突地联合起来，减少人工调参的难度，实现端到端的快速训练有进一步的研究价值。

（4）边缘端AI芯片限制条件下的应用

民用方面如在手机、智能监控摄像头以及移动穿戴式设备对神经网络能够部署到边缘端智能芯片的应用日益增多，军用方面如无人机敌情侦察、卫星导弹的智能化也有着同样迫切的需求。但是在面对边缘端AI芯片的算力和功耗等限制条件下，知识蒸馏技术压缩得到的小型网络参数的精度仍然很高，会消耗非常多的计算资源，因此知识蒸馏技术结合参数量化技术对神经网络从学术研究到工业落地有着巨大的应用潜力，需要结合特定的应用场景、计算环境约束、待分析的目标特性、特殊的知识类型，展开个性化研究。

（5）评价标准规范化

目前对网络压缩中的知识蒸馏技术的评价主要侧重于准确率、模型内存等方面，但使用更加全面的评价指标对于发现不同算法的优缺点是大有裨益的。

由于在工程上经常有快速训练的需求，不同算法训练时消耗的计算资源和训练时间也可以作为一种衡量算法的评价指标；由于简化后的网络模型往往是在资源受限的设备上运行，乘加运算量、硬件能耗也是重要的评价指标。

由于知识蒸馏压缩的效果与教师、学生网络的选取以及压缩的程度有密切的关系，在超出一定压缩程度后，某些知识蒸馏方法的压缩效果会急剧下降。因此，在知识蒸馏的理论研究中，非常有必要建立标准完备的用于压缩效果评价的网络集合。该集合包含多个网络压缩对，每个网络压缩对都单独对应一个教师网络和一个学生网络（如网络压缩对ResNet152与ResNet101，ResNet152与ResNet50，WRN40-2与WRN40-1，WRN40-2与WRN16-2等），每个网络压缩对也都标志着具体的网络结构和压缩程度。网络压缩中各种知识蒸馏算法都使用这个集合在固定的公开数据集上进行蒸馏效果的比较，以呈现该算法在不同网络结构、不同压缩程度下的压缩效果，以使实验结果更加具有说服力，以更清晰地呈现不同算法的优劣性和泛化性。

4 结束语

本文简述了网络压缩方法的发展渊源，在简单对比典型网络压缩技术的基础上，重点针对近年深度神经网络压缩知识蒸馏技术进行了详细的梳理，全面介绍了近年知识蒸馏技术的典型探索，并给出了未来该技术研究拓展的初步思考，希望对当前及未来知识蒸馏技术的研究工作有所帮助。

参考文献：

[1] JIANG Z T, QIN J Q, ZHANG S Q. Parameterized pooling convolution neural network for image classification[J]. Acta Electronica Sinica, 2020, 48(9): 1729-1734.
江泽涛, 秦嘉奇, 张少钦. 参数池化卷积神经网络图像分类方法[J]. 电子学报, 2020, 48(9): 1729-1734.

[2] LIU Y, ZHAN Y W. Survey of small object detection algorithms based on deep learning[J]. Computer Engineering and Applications, 2021, 57(2): 37-48.
刘洋, 战荫伟. 基于深度学习的小目标检测算法综述[J]. 计算机工程与应用, 2021, 57(2): 37-48.

[3] TIAN Q C, MENG Y. Image semantic segmentation based on convolutional neural network[J]. Journal of Chinese Computer Systems, 2020, 41(6): 1302-1313.
田启川, 孟颖. 卷积神经网络图像语义分割技术[J]. 小型微型计算机系统, 2020, 41(6): 1302-1313.

[4] FU Z Y, ZHOU S J, LI D G. Lightweight target recognition deep neural network and its application[J]. Computer Engineering and Applications, 2020, 56(18): 131-136.
付佐毅, 周世杰, 李顶根. 轻量级目标识别深度神经网络及其应用[J]. 计算机工程与应用, 2020, 56(18): 131-136.

[5] WANG Y C, LI Z H, HAO H Y, et al. Research on visual perception technology of autonomous driving based on improved convolutional neural network[J]. Journal of Physics: Conference Series, 2020, 1550: 032103.

[6] KULIKA S, SHTANKOA A. Using convolutional neural networks for recognition of objects varied in appearance in computer vision for intellectual robots[J]. Procedia Computer Science, 2020, 169: 164-167.

[7] KU H C, DONG W. Face recognition based on MTCNN and convolutional neural network[J]. Frontiers in Signal Processing, 2020, 4(1): 37-42.

[8] HAMEED N, SHABUT A, HAMEED F, et al. Mobile-based skin lesions classification using convolution neural network [J]. Annals of Emerging Technologies in Computing, 2020, 4

(2): 1-12.

[9] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 770-778.

[10] LECUN Y, DENKER J S, SOLLA S A. Optimal brain damage[C]//Proceedings of the Advances in Neural Information Processing Systems, Denver, Nov 27-30, 1989. San Mateo: Morgan Kaufmann, 1989: 598-605.

[11] CARREIRA-PERPIÑÁN M Á, IDELBAYEV Y. "Learning-compression" algorithms for neural net pruning[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 8532-8541.

[12] WEN W, WU C P, WANG Y D, et al. Learning structured sparsity in deep neural networks[C]//Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Dec 5-10, 2016. Red Hook: Curran Associates, 2016: 2074-2082.

[13] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets[C]//Proceedings of the 5th International Conference on Learning Representations, Toulon, Apr 24-26, 2017: 1-13.

[14] COURBARIAUX M, BENGIO Y, DAVID J P. BinaryConnect: training deep neural networks with binary weights during propagations[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2015, Montreal, Dec 7-12, 2015. Red Hook: Curran Associates, 2015: 3123-3131.

[15] LI F, ZHANG B, LIU B. Ternary weight networks[C]//Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Dec 5-10, 2016. Red Hook: Curran Associates, 2016: 2024-2032.

[16] XU Y H, WANG Y Z, ZHOU A J, et al. Deep neural network compression with single and multiple level quantization[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, New Orleans, Feb 2-7, 2018. Menlo Park: AAAI, 2018: 4335-4342.

[17] LIN Z H, COURBARIAUX M, MEMISEVIC R, et al. Neural networks with few multiplications[C]//Proceedings of the 4th International Conference on Learning Representations, San Juan, May 2-4, 2016: 1-9.

[18] JADERBERG M, VEDALDI A, ZISSERMAN A. Speeding up convolutional neural networks with low rank expansions [J]. arXiv:1405.3866, 2014.

[19] LEBEDEV V, GANIN Y, RAKHUBA M, et al. Speeding-up convolutional neural networks using fine-tuned CP-decomposition [C]//Proceedings of the 3rd International Conference on Learning Representations, San Diego, May 7-9, 2015: 1-11.

[20] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. arXiv:1704.04861, 2017.

[21] HUANG G, SUN Y, LIU Z, et al. Deep networks with stochastic depth[C]//LNCS 9908: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, Oct 11-14, 2016. Cham: Springer, 2016: 646-661.

[22] KIM J, PARK Y, KIM G, et al. SplitNet: learning to semantically split deep networks for parameter reduction and model parallelization[C]//Proceedings of the 34th International Conference on Machine Learning, Sydney, Aug 6-11, 2017. New York: ACM, 2017: 1866-1874.

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

[24] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer[C]//Proceedings of the 5th International Conference on Learning Representations, Toulon, Apr 24-26, 2017: 1-13.

[25] SRINIVAS S, FLEURET F. Knowledge transfer with Jacobian matching[C]//Proceedings of the 35th International Conference on Machine Learning, Stockholm, Jul 10-15, 2018. New York: ACM, 2018: 4730-4738.

[26] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. arXiv:1406.2661, 2014.

[27] LANG L, XIA Y Q. Survey on compact neural network model design[J]. Journal of Frontiers of Computer Science and Technology, 2020, 14(9): 1456-1470.
郎磊, 夏应清. 紧凑的神经网络模型设计研究综述[J]. 计算机科学与探索, 2020, 14(9): 1456-1470.

[28] JI R R, LIN S H, CHAO F, et al. Deep neural network compression and acceleration: a review[J]. Journal of Computer Research and Development, 2018, 55(9): 1871-1888.
纪荣嵘, 林绍辉, 晁飞, 等. 深度神经网络压缩与加速综述[J]. 计算机研究与发展, 2018, 55(9): 1871-1888.

[29] LIN J D, WU X Y, CHAI Y, et al. Structure optimization of convolutional neural networks: a survey[J]. Acta Automatica Sinica, 2020, 46(1): 24-37.

林景栋, 吴欣怡, 柴毅, 等. 卷积神经网络结构优化综述[J]. 自动化学报, 2020, 46(1): 24-37.

[30] GENG L L, NIU B N. Survey of deep neural networks model compression[J]. Journal of Frontiers of Computer Science and Technology, 2020, 14(9): 1441-1455.
耿丽丽, 牛保宁. 深度神经网络模型压缩综述[J]. 计算机科学与探索, 2020, 14(9): 1441-1455.

[31] GAO H, TIAN Y L, XU F Y, et al. Survey of deep learning model compression and acceleration[J]. Journal of Software, 2021, 32(1): 68-92.
高晗, 田育龙, 许封元, 等. 深度学习模型压缩与加速综述[J]. 软件学报, 2021, 32(1): 68-92.

[32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the 3rd International Conference on Learning Representations, San Diego, May 7-9, 2015: 1-14.

[33] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[34] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, Dec 3-6, 2012. Red Hook: Curran Associates, 2012: 1097-1105.

[35] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//LNCS 8689: Proceedings of the 13th European Conference on Computer Vision, Zurich, Sep 5-12, 2014. Cham: Springer, 2014: 818-833.

[36] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, Jun 7-12, 2015. Washington: IEEE Computer Society, 2015: 1-9.

[37] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 2261-2269.

[38] CHEN Y P, LI J N, XIAO H X, et al. Dual path networks [C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2017, Los Angeles, Dec 4-9, 2017. Red Hook: Curran Associates, 2017: 4467-4475.

[39] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-23, 2018. Piscataway: IEEE, 2018: 1-13.

[40] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 7341-7349.

[41] SANDLER M, HOWARD A G, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 4510-4520.

[42] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[J]. arXiv:1602.07360, 2016.

[43] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 6848-6856.

[44] LI Q Q, JIN S Y, YAN J J. Mimicking very efficient network for object detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 7341-7349.

[45] WANG T, YUAN L, ZHANG X P, et al. Distilling object detectors with fine-grained feature imitation[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 4933-4942.

[46] ZHANG T T, DONG J Y, ZHAO H R, et al. Lightweight phytoplankton detection network based on knowledge distillation[J]. Journal of Applied Sciences, 2020, 38(3): 367-376.
张彤彤, 董军宇, 赵浩然, 等. 基于知识蒸馏的轻量型浮游植物检测网络[J]. 应用科学学报, 2020, 38(3): 367-376.

[47] HOU Y N, MA Z, LIU C X, et al. Inter-region affinity distillation for road marking segmentation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 13-19, 2020. Piscataway: IEEE, 2020: 12483-12492.

[48] SHAO H K, ZHONG D X, DU X F. Towards efficient unconstrained palmprint recognition via deep distillation Hashing[J]. arXiv:2004.03303, 2020.

[49] BHARDWAJ S, SRINIVASAN M, KHAPRA M M. Efficient video classification using fewer frames[C]//Proceedings

of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 354-363.

[50] WU H Y, LIU J, XIE Y, et al. Knowledge transfer dehazing network for nonhomogeneous dehazing[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. Piscataway: IEEE, 2020: 1975-1983.

[51] YANG C L, XIE L X, QIAO S Y, et al. Training deep neural networks in generations: a more tolerant teacher educates better students[C]//Proceedings of the 2019 AAAI Conference on Artificial Intelligence, Hawaii, Jan 27-Feb 1, 2019. Menlo Park: AAAI, 2019: 5628-5635.

[52] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 2818-2826.

[53] PEREYRA G, TUCKER G, CHOROWSKI J, et al. Regularizing neural networks by penalizing confident output distributions[C]//Proceedings of the 5th International Conference on Learning Representations, Toulon, Apr 24-26, 2017: 1-12.

[54] CHO J H, HARIHARAN B. On the efficacy of knowledge distillation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 4793-4801.

[55] ROMERO A, BALLAS N, KAHOU S E, et al. FitNets: hints for thin deep nets[C]//Proceedings of the 3rd International Conference on Learning Representations, San Diego, May 7-9, 2015: 1-13.

[56] LEE S H, KIM D H, SONG B C. Self-supervised knowledge distillation using singular value decomposition[C]//LNCS 11210: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 339-354.

[57] YIM J, JOO D, BAE J H, et al. A gift from knowledge distillation: fast optimization, network minimization and transfer learning[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 7130-7138.

[58] HUANG Z H, WANG N Y. Like what you like: knowledge distill via neuron selectivity transfer[J]. arXiv:1707.01219, 2017.

[59] AHN S, HU S X, DAMIANOU A C, et al. Variational information distillation for knowledge transfer[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 9163-9171.

[60] KIM J, PARK S, KWAK N. Paraphrasing complex network: network compression via factor transfer[J]. arXiv:1802.04977, 2018.

[61] TIAN Y L, KRISHNAN D, ISOLA P. Contrastive representation distillation[C]//Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Apr 26-30, 2020: 1-15.

[62] ZHANG Z, NING G H, HE Z H. Knowledge projection for deep neural networks[J]. arXiv:1710.09505, 2017.

[63] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 3967-3976.

[64] STOCK P, JOULIN A, GRIBONVAL R, et al. And the bit goes down: revisiting the quantization of neural networks [C]//Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Apr 26-30, 2020: 1-11.

[65] LIU Y F, CAO J J, LI B, et al. Knowledge distillation via instance relationship graph[C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. Piscataway: IEEE, 2019: 7096-7104.

[66] GRIPON V, ORTEGA A, GIRAULT B. An inside look at deep neural networks using graph signal processing[C]//Proceedings of the 2018 Information Theory and Applications Workshop, San Diego, Feb 11-16, 2018. Piscataway: IEEE, 2018: 1-9.

[67] ANIRUDH R, THIAGARAJAN J J, SRIDHAR R, et al. MARGIN: uncovering deep neural networks using graph signal analysis[J]. arXiv:1711.05407, 2017.

[68] LASSANCE C, GRIPON V, ORTEGA A. Laplacian networks: bounding indicator function smoothness for neural network robustness[J]. APSIPA Transactions on Signal and Information Processing, 2021, 10: 1-12.

[69] LASSANCE C, BONTONOU M, HACENE G B, et al. Deep geometric knowledge distillation with graphs[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, May 4-8, 2020. Piscataway: IEEE, 2020: 8484-8488.

[70] FENG Y S, WANG H, HU R, et al. Triplet distillation for deep face recognition[C]//Proceedings of the 2020 IEEE International Conference on Image Processing, Abu Dhabi, Oct 25-28, 2020. Piscataway: IEEE, 2020: 808-812.

[71] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv:1703.07737, 2017.

[72] TUNG F, MORI G. Similarity-preserving knowledge distillation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 1365-1374.

[73] PENG B Y, JIN X, LI D S, et al. Correlation congruence for knowledge distillation[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 5006-5015.

[74] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Dec 7-13, 2015. Washington: IEEE Computer Society, 2015: 1449-1457.

[75] LEE S, SONG B C. Graph-based knowledge distillation by multi-head attention network[J]. arXiv:1907.02226, 2019.

[76] YADAV A K, SHAH S, XU Z, et al. Stabilizing adversarial nets with prediction methods[J]. arXiv:1705.07364, 2017.

[77] WANG X J, ZHANG R, SUN Y, et al. KDGAN: knowledge distillation with generative adversarial networks[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2018, Montreal, Dec 3-8, 2018. Red Hook: Curran Associates, 2018: 783-794.

[78] GAO D, ZHUO C. Private knowledge transfer via model distillation with generative adversarial networks[J]. arXiv: 2004.04631, 2020.

[79] XU Z, HSU Y C, HUANG J W. Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks[J]. arXiv:1709.00513, 2017.

孟宪法（1997—），男，山东临沂人，硕士研究生，主要研究方向为深度学习、图像处理。

MENG Xianfa, born in 1997, M.S. candidate. His research interests include deep learning and image processing.

刘方（1970—），女，山东济南人，博士，副教授，主要研究方向为信息融合、图像分析。

LIU Fang, born in 1970, Ph.D., associate professor. Her research interests include information fusion and image analysis.

李广（1995—），男，湖南娄底人，硕士研究生，主要研究方向为深度学习、网络压缩剪枝。

LI Guang, born in 1995, M.S. candidate. His research interests include deep learning and network compression pruning.

黄萌萌（1998—），女，安徽马鞍山人，硕士研究生，主要研究方向为深度学习、遥感图像目标识别。

HUANG Mengmeng, born in 1998, M.S. candidate. Her research interests include deep learning and remote sensing image target recognition.