中图法分类号:TP37　　**文献标识码**:A　　**文章编号**:1006-8961(2023)04-0935-28

论文引用格式:Sun R Y and Xiong H K. 2023. Model distillation for high-level semantic understanding: a survey. Journal of Image and Graphics, 28(04):0935-0962(孙若禹, 熊红凯. 2023. 高层语义分析中的模型蒸馏方法综述. 中国图象图形学报, 28(04):0935-0962)[DOI:10.11834/jig.210337]

高层语义分析中的模型蒸馏方法综述

孙若禹,熊红凯*

上海交通大学电子工程系,上海　200240

摘　要:计算机视觉的任务目标是建立接近人类视觉系统的计算模型。随着深度神经网络(deep neural network,DNN)的发展,对计算机视觉中高层语义的分析与理解成为研究重点。计算机视觉的高层语义通常为人类可理解、可表述的用于表达图像、视频等媒体信号内容的描述子(descriptor),典型的高层语义分析任务包含图像分类、目标检测、实例分割、语义分割与视频场景识别、目标跟踪等。基于深度神经网络的算法使计算机视觉任务获得逐步提升的性能,但是网络模型的体量增大与计算效率的降低随之而来。模型蒸馏是一种基于迁移学习进行模型压缩的方案。此类方案通常利用一个预训练模型作为教师,提取其有效的表示,如模型输出、隐藏层特征或特征间相似度等,并将上述表示作为另一个规模较小、推断速度较快的学生模型的额外监督信号,对该学生模型进行训练,以达到提升小模型性能从而取代大模型的目的。模型蒸馏对模型性能与计算复杂度有着良好权衡,因此愈来愈多地用于基于深度学习的高层语义分析中。自2014年模型蒸馏概念提出以来,研究人员开发了大量应用于高层语义分析的模型蒸馏方法,在图像分类、目标检测与语义分割任务中的应用最为广泛。本文对上述典型任务中具有代表性的模型蒸馏方案进行调研和汇总,依照不同的视觉任务进行介绍。首先,从最成熟、应用最广泛的分类任务模型蒸馏方法开始,介绍其不同的设计思路与应用场景,展示部分实验性能的对比,指出在分类任务上与在检测、分割任务上应用模型蒸馏的条件差异性。接着,对几种经特殊设计而应用于目标检测、语义分割的典型模型蒸馏方法进行介绍,结合模型结构对设计目的与思路进行说明,提供部分实验结果的对比与分析。最后,对当前高层语义分析中模型蒸馏方法的现状进行了总结分析,并指出存在的困难及不足,设想未来可能的探索思路与发展方向。

关键词:模型蒸馏;深度学习;图像分类;目标检测;语义分割;迁移学习

Model distillation for high-level semantic understanding: a survey

Sun Ruoyu, Xiong Hongkai*

Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai　200240, China

Abstract: Computer vision tasks aim to construct computational models in relevant to functions-like of human visual systems. Current deep learning models are progressively improving upper bounds of performances in multiple computer vision tasks, especially for analysis and understanding of high-level semantics, i. e., multimedia-based descriptors for human recognition. Typical tasks to understand high-level semantics include image classification, object detection, instance segmentation, semantic segmentation, and video's recognition and tracking. With the development of convolutional neural networks (CNNs), deep learning based high-level semantic understanding have all been benefiting from increasingly deeper and cumbersome models, which is also challenged for the problem of storages and computational costs. To obtain

收稿日期:2021-05-18;**修回日期**:2022-03-05;**预印本日期**:2022-03-12

*通信作者:熊红凯　xionghongkai@sjtu.edu.cn

基金项目:国家自然科学基金重点项目(61932022)

Supported by:National Natural Science Foundation of China (61932022)

lighter structure and computation efficiency, many model compression strategies have been proposed, e. g. , pruning, weight quantization, and low-rank factorization. But, such challenging issue is to be resolved for altered network structure or drop-severe of performance when deployed on computer vision tasks. Model distillation can be as one of the typical compression methods in terms of transfer learning to model compression. In general, model distillation utilizes a large and complicated pre-trained model as "teacher" and takes its effective representations, e. g. , model outputs, features of hidden layers or feature maps-between similarities. These representations are treated as extra supervision signal together with the original ground truth for a lighter and faster model's training, in which the lighter model is called "student". As model distillation provides favorable balance between models' performances and efficiency, it is being rapidly explored on different computer vision tasks. This paper investigates the progress of model distillation methods since its introduction in 2014 and introduces their different strategies in various applications. We review some popular distillation strategies and current model distillation algorithms deployed on image classification, object detection and semantic segmentation in this paper. First, we introduce distillation methods for image classification tasks, where model distillation has already achieved mature development. Fundamentals of model distillation starts from using teacher classifiers' output logits as soft labels, bringing student with more inter-categories structural information, which is not available in conventional one-hot ground truths. Furthermore, hint learning can be used to utilize hierarchical structure of neural networks and take feature maps from hidden layers as another "teachers"-involved representations. Most of distillation strategies are designed and derived from similar approaches. In the aspects of frameworks' design and application scenes, the paper respectively introduced some typical distillation strategies on classification models. Some methods mainly considered novel approaches on supervision signal design, i. e. , ensembles that differs from conventional classification soft labels or feature maps. Newly developed features for student models to mimic are usually computed from attention or similarity maps of different layers, data augmentations or sampled images. Other methods consider adding noise or perturbation to teacher classifiers' output or using probability inference to minimize the gap between teacher and student models. These specially designed features or logits are focused on a more appropriate representation of knowledge in teacher models than plain features from some layers' outputs. Moreover, in other methods, the procedure of model distillation is altered, and more complicated schemes are introduced to transfer teacher's knowledge instead of simply training the student with generated labels or features. Also, as generative adversarial networks (GANs) achieve promising performance in image synthesis, some model distillation methods also introduce adversarial mechanisms in classifiers' distillation, where teacher models' features are regarded as "real ones" and the students are expected to "generate" similar features. In many practical scenes such as model compression, self-training and parallel computing, classifiers' distillation is utilized in coordinate to specific process as well, e. g. , fine tuning networks with full-precision teachers, distilling student model with its previous versions during training, and using models from different nodes as teachers. We summarize some popular strategies performances and illustrate the data in a table after approaches of model distillation in image classification tasks are introduced. Distillation methods' performances on improving classifiers' top-1 accuracies are compared on several typical classification datasets. The second part of the paper focuses on specially developed distillation methods for computer vision tasks more complicated than classification, e. g. , object detection, instance segmentation and semantic segmentation. Differentiated from classifiers, models of these tasks contain more redundant structures with heterogeneous outputs. Hence, recent works on detectors' and segmentation models' distillation is relatively less than those in classifiers' distillation. The paper describes current challenges in designing of distillation frameworks on detection and segmentation tasks. Some of typical distillation methods for detectors and segmentation models are then introduced based on different tasks and their multifaceted structures. Since there were few works specified for instance segmentation models' distillation, the papers simply introduce similar distillation methods for object detectors in the beginning of the second part. For detectors, requirements from localization demand special concentration on local information around foreground objects. Meanwhile, images from object detection datasets consists of more complicated scenes generally in which large amounts of different objects may occur. Hence, the solutions of distillation strategies-borrowing from for classifies may bring undesired performance decrease in object detection. Due to more complex structures in detectors, previous distillation methods may not be applicable. As "backbone with task heads" structure is widely used in modern computer vision models, researchers develop novel distillation methods mainly based on this typical framework.

The introduced detectors' distillation strategies investigate issues above and mainly focus on specific output logits acquirement and specially designed loss functions for different parts in detectors. To highlight foreground regions before distillation, backbones-derived feature maps are often selected through regions of interest (RoIs) using masking operations. Various of output logits are selected in different methods from teacher models' task heads, affecting training of students' task heads in terms of specific matching and imitation schemes. Semantic segmentation requires more global information than object detection or instance segmentation tasks, focusing on pixel-wise classification inside the total image. One of the critical factors of pixels' correct classification is oriented to the analysis of inter-pixel relationships. Hence, model distillation methods for semantic segmentation also take advantages of pixels in both output masks and feature maps from hidden layers. Distillation strategies introduced in the paper are majorly on the application of hierarchical distillation on different part, e. g. , the imitation of full output classification mask, imitation of full feature maps, computing of similarity matrices, and using conditional GANs (cGANs) for auxiliary imitation. The former two approaches are fundamental practices in model distillation. In contrast, to realize segmentation model's pixel-wise knowledge to be more 'compact' after compression, some distillation methods utilize compressed features instead of original one to compute similarity with student. When cGANs is used to imitate student segmentation model to the teacher features, researchers introduce Wasserstein distance as a better metric for adversarial training. At the final part of this paper, previous works of model distillation for high-level semantic understanding are summarized. We review some obstacles and unsolved problems in current development of model distillation, and the future research direction is predicted as well.

Key words: model distillation; deep learning; image classification; object detection; semantic segmentation; transfer learning

0 引 言

近年来，深度神经网络(deep neural network，DNN)取得了迅猛发展，通过数据与标签对神经网络进行训练，各类实际应用任务的性能获得了大幅提升。计算机视觉作为一种实际应用极为广泛的任务，旨在对人类的视觉系统建立计算模型，或是构建能够模拟人类视觉功能的自动化系统(Huang，1996)。图像、视频等媒体通常具备能够被人类理解、被语言描述的结构化信息，例如表示媒体类别、包含物体和相似度等内容的关键词或其他描述子(descriptor)，这些信息称为高层语义(Huang，1996；Jiang 等，2005；Liu 等，2007)，对高层语义特征的提取、分析与理解是计算机视觉的重要任务(Wu，2007)。例如，图像分类、目标检测和语义分割等任务属于典型的高层语义分析(Shapiro，2019)。得益于深度神经网络，尤其是卷积神经网络(convolutional neural network，CNN)的特征提取作用，高层语义分析获得了愈来愈高的性能。由于经过训练的深度网络往往具备对输入信号进行高层特征提取的能力，其对大量计算机视觉任务提供了显著的帮助(Sinha 等，2018；Voulodimos 等，2018；Feng 等，2019)。然而，对基于深度模型的算法而言，更高的性能通常意味着对模型体量的更大需求。目前，典型的高层语义分析任务，如图像分类、目标检测和语义分割等任务中，能够大幅提升性能的算法通常依赖于更加复杂的模型。对用于特征提取的主干网络(backbone)而言，深度卷积网络层数不断加深，旁路分支逐渐复杂，例如 ResNet(He 等，2016)系列从 ResNet-18、ResNet-50 加深至 ResNet-152，Inception(Szegedy 等，2015)系列网络从 V1 扩展至 V4，近年来更是出现了 ResNeXt(Xie 等，2017)和 ResNeSt(Zhang 等，2022)等结构更为复杂、运算更加繁多的主干网络，用于提取更丰富的语义信息；同时，可变形卷积(Dai 等，2017)和多尺度空洞卷积(Yu 和 Koltun，2016)等操作的引入使计算复杂度进一步提升；针对不同任务的具体模型，例如检测器、分割器等模型的多任务分支也较分类器更为复杂，为任务所需的计算资源与存储空间引入巨大负担。因此，对视觉模型进行压缩成为一项重要任务。

对神经网络参数进行压缩的方案，例如网络剪枝(Han 等，2015；Park 等，2017；Tung 和 Mori，2018)、权重量化(Wu 等，2016；Zhou 等，2017；Han 等，2016)等，通常可将网络规模压缩至极小的量级，然而一般伴随着模型性能的显著下降。而计算机视觉任务

中，性能是最重要的指标之一，这便要求研究人员设计的模型压缩方案能够同时保证较高性能与模型轻量化。

Hinton等人（2015）提出的知识蒸馏概念，其自迁移学习的概念引入，将单个或多个预训练、结构复杂而性能较高的教师模型输出作为额外的软标签，与ground truth标签共同对一个体量较小而性能较差的分类网络即学生模型进行训练，以此将教师模型中的知识传递至学生模型，使其获得常规训练中无法获取的性能表现，并称这种训练方式为蒸馏。通过上述算法，在训练完成后，利用轻量学生模型替换教师模型，便可削弱模型规模与推断速度的劣势。同时，得益于蒸馏过程对学生模型性能的额外提升，学生模型的最终性能距教师模型性能更为接近。因此，学生模型替代教师模型后，相较于后者的性能损失通常能够控制在可接受的范围内。

事实上，上述蒸馏方案构成了对轻量级模型的一种额外性能提升方法。一般地，此类基础的知识蒸馏方式称为模型蒸馏，图1展示了以常见的分类网络为例面向常规分类网络的通用模型蒸馏框架。研究人员在此基础上提出了利用数据增强进行的数据蒸馏（Radosavovic等，2018）以及利用不同数据集进行的任务蒸馏（Girdhar等，2019），使知识蒸馏在迁移学习与半监督学习任务中获得了推广。然而，针对计算机视觉任务，应用最广泛、形式最多样的知识蒸馏方式仍然是模型蒸馏。

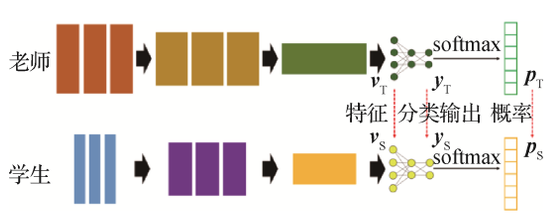

图1　一般的模型蒸馏示意图

Fig. 1　General schematic diagram of model distillation

针对高层语义的分析与理解中，图像分类对深度模型的应用最为成熟，模型蒸馏算法在此任务上取得了长足发展。如图1所示，教师模型有3种典型的监督方式，即使用分类概率向量作为软标签、利用分类器输出的结果向量作为监督信号和提取教师模型中间层输出特征图进行线索学习（Romero等，2015）。以此为基础，各类模型蒸馏算法不断革新分类网络的蒸馏方式。针对分类模型蒸馏的问题与难点，主要基于两类方向进行研究。1）针对分类模型蒸馏方法本身，对模型蒸馏的策略、算法与损失函数进行更新颖的设计。例如，引入噪声，以平衡教师、学生的分布差异；面对性能差距，将蒸馏过程细化，引入逐层或多步骤蒸馏方案；更精细地设计线索学习的方法，缓解特征直接逼近带来的优化困难；使用特征、输出张量计算层间的、样本间的相关关系，并以此关系作为进行逼近的信号，获得更灵活的知识传递方法；或是回避利用损失函数进行直接逼近的蒸馏方案，引入对抗判决器等方法，隐式地使学生学习教师模型的输出。2）由于以分类模型为基础的部分特殊任务需要额外的模型性能提升，研究人员考察了模型蒸馏在特殊任务中的应用方法。例如，利用自蒸馏以改良分类器的训练方式；在域迁移、人脸识别等特殊任务中蒸馏特定任务的模型；与模型量化、剪枝等模型压缩方法并行，提升轻量网络性能；以及在分布式、加速训练的场景下考察多模型间的蒸馏。

由于目前对模型蒸馏的研究仍主要集中在分类器蒸馏，本文在第2节对以上两类方向（新颖模型蒸馏策略设计、模型蒸馏在特定任务中的应用）进行介绍。

针对较图像分类更复杂的高层语义分析，例如目标检测、实例分割及语义分割等，研究人员提出了相应的模型蒸馏策略。由于与分类模型的蒸馏具有显著不同，面临的难点与提出的相应方案主要包含以下方面：针对检测、分割模型结构更复杂、输出更多样化的特点，对检测器、分割器考察了不同网络分支的作用，面向不同分支分别设计监督方案，例如损失函数设计；针对目标检测任务关注各目标附近局部区域、语义分割任务考察逐像素分类的特点，考察了不同局部信息携带的知识，并对不同层、不同位置的特征应用独特蒸馏方式。

由于面向检测、分割等任务的模型蒸馏研究仍处于发展阶段，有关研究相对较少，本文在第3节选取典型的若干方案进行介绍，并尽可能地在相同实验配置下，汇总不同模型蒸馏方法的实验性能，以进行比较与分析。

不同的模型蒸馏方法对模型输出、损失函数及输入样本的选择均有所不同，本文介绍模型蒸馏工作时涉及了公式表述。为行文规范，本文将表示相同种类变量的公式符号进行统一，如表1所示。其

中包含了本文多数的统一符号表示，部分特殊变量类型不使用以上符号，未列于表中。对于同种类别但具体含义不同的变量，利用不同上标与下标进行区分，且各符号的具体含义在正文进行了解释。

表 1　本文涉及的主要公式符号及含义

Table 1　Definitions of common symbols in this paper's formula

符号	含义	符号	含义
$f(\cdot)$	泛指神经网络模型	L	泛指损失函数
$\boldsymbol{I}$	输入图像	$\boldsymbol{B}$	检测框
$\boldsymbol{v}$	各类特征图	$\boldsymbol{y}$	分类器输出向量
$\boldsymbol{p}$	各类概率向量	$\boldsymbol{M}$	掩膜类型的张量
$\boldsymbol{w}$	模型权重参数	p,q	单个概率值或置信度
ϵ	阈值	y	分类标签
ξ	松弛阈值	$\boldsymbol{\xi}$	噪声扰动
$I(\cdot;\cdot)$	互信息	$r(\cdot)$	回归层/后处理层
M,N	样本数、目标数、类别数等	C,H,W	特征的通道数、高、宽
$\boldsymbol{P}$	候选提取区域	$\varphi(\cdot,\cdot)$	泛指核函数
λ	权重/平衡因子	$\boldsymbol{r}$	回归器输出
w	单个权重参数	u,v	特征图中的单个像素
$\boldsymbol{A}$	泛指各类矩阵	$\boldsymbol{R}$	（特征图中的）区域

总体上，本文调研了自2014年知识蒸馏概念提出后，研究人员针对典型的高层语义分析任务——分类、检测和分割等提出的模型蒸馏方法，对近年来模型蒸馏的各种方案进行了介绍、对比与分析，并对未来模型蒸馏针对计算机视觉任务的设计方案进行了展望。本文主要内容分布于第2节和第3节，其内容结构安排如图2所示，分别对3类视觉任务的模型蒸馏进行介绍。

本文结构说明如下：第1节介绍分类、检测和分割任务中常用的性能评估标准、数据集以及模型压缩中的部分评估指标；第2节对较成熟的分类任务中的模型蒸馏方案进行介绍，展示对比不同分类器蒸馏方法使模型获得的性能提升；第3节对适用于检测、分割任务的模型蒸馏方法及针对不同模型特征的处理方式进行介绍，分析基本的设计思路并进行性能对比；第4节总结上述各类工作的特点与存在的不足，在此基础上对计算机视觉任务中模型蒸馏方法的未来发展趋势进行展望。

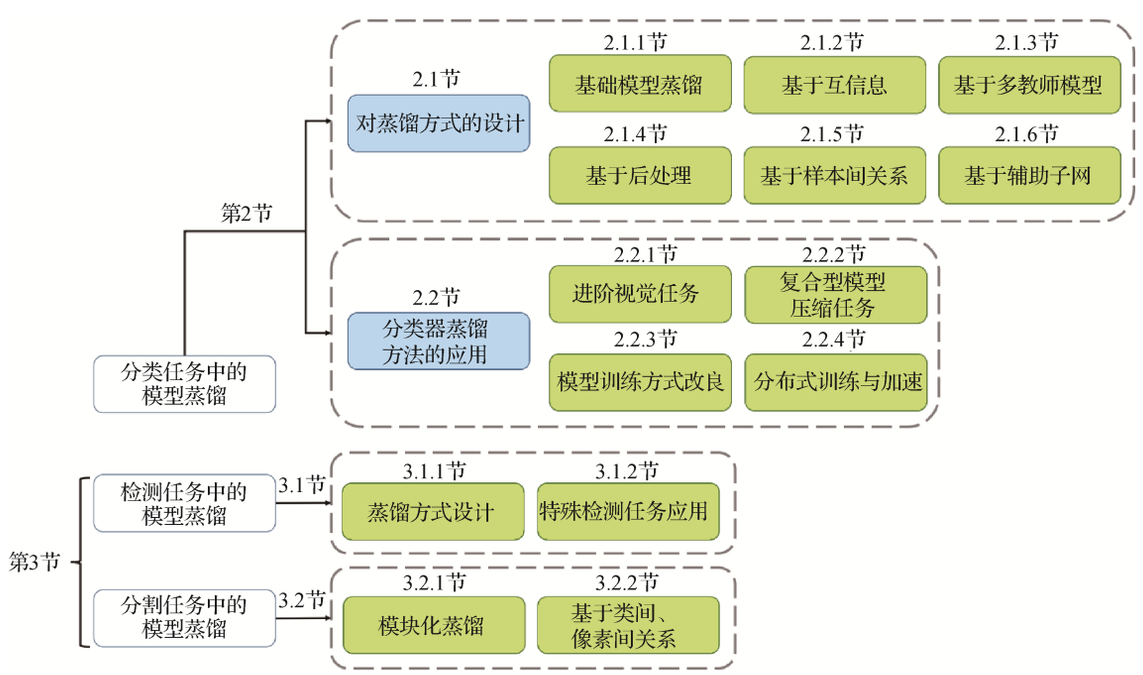

图2　本文第2节和第3节的内容安排

Fig. 2　Arrangements of section 2 and section3 in this paper

1　高层语义分析与模型蒸馏概念

1.1　高层语义分析及其性能评价指标

在对媒体高层语义信息进行提取与理解的计算机视觉任务中，图像分类是最典型的任务之一，也是应用深度神经网络最为广泛的视觉任务。一般形式为：给定输入图像$\boldsymbol{I}$，分类器$f_c(\cdot)$提供如下输出

$$\boldsymbol{p}=f_c(\boldsymbol{I}) \tag{1}$$

式中，$\boldsymbol{p}$为N_{cls}维分类概率向量，向量各个维度的元

素代表分类器认为图像$\boldsymbol{I}$属于N_{cls}类中某一类的概率(置信度)。对分类标注为$\boldsymbol{y}$的图像,其ground truth一般表示为one-hot向量$\boldsymbol{p}_{\text{gt}}$,其中仅对应$\boldsymbol{y}$类别的位置元素置为1,其余位置元素置零。进行损失函数计算时,通常使用交叉熵损失函数,具体为

$$L_{\text{CE}} = \boldsymbol{p}_{\text{gt}}^{\text{T}} \log \boldsymbol{p} \tag{2}$$

当且仅当$\boldsymbol{p}_{\text{gt}} = \boldsymbol{p}$时,损失$L_{\text{CE}} = 0$。

模型训练完成后,在推断阶段选取模型输出的概率向量,将概率按降序排列以进行评估。当前通用的评估指标为Top-1准确率、Top-5准确率、Top-1错误率和Top-5错误率。前两者与后两者是完全相关的。首先,对任意一幅测试集图像,若输出向量中置信度(概率)最高(Top-1)的元素所对应的类别的确为该图像所属类别,则该图像属于正确分类。Top-1准确率表示测试集中所有分类结果正确的图像占全部测试集图像的比率,错误率则表示剩余分类错误的图像的占比;类似地,若在降序排列的前5个置信度中,包含有图像所属的实际类别,则认为此为Top-5意义下的正确分类样本,其准确率与错误率的计算方式与Top-1意义下的计算方式一致。

与图像分类不同,目标检测任务通常表示为

$$\boldsymbol{B} = f_{\text{OD}}(\boldsymbol{I}) \tag{3}$$

式中,$\boldsymbol{B}$为由若干检测框的位置、大小以及属于相应类别的置信度向量组成的向量组。目标检测要求检测器$f_{\text{OD}}(\cdot)$对输入图像$\boldsymbol{I}$中可定位与计数的前景物体进行外接矩形框标注,而前沿的检测任务同时要求判断各物体所属类别。因此,目标定位任务已被目标检测完全取代。

实例分割任务较目标检测更进一步,要求对图像中属于各目标物体的像素点进行分割,标明前景像素点所属类别与实例通常以掩膜表示。不同模型对目标检测与实例分割存在不同处理方式。基于锚框的检测器,如单阶段目标检测器YOLO(you only look once)(Redmon等,2016)系列与两阶段检测器Faster R-CNN(region CNN)(Ren等,2017)均将目标检测任务视做目标框分类与回归任务;无须锚框的检测器如CenterNet(Zhou等,2019)则将任务转化为对目标特定关键点的定位与目标框大小的回归问题。对于实例分割任务,Mask R-CNN(He等,2017)利用检测结果与全卷积子网络共同辅助分割,Instance-sensitive FCN(fully convolutional network)(Dai等,2016a)则对像素在各实例中的相对位置进行编码,SOLO(segmenting objects by locations)(Wang等,2020b)从物体中心点出发,分别预测实例的类别与掩膜。总之,不同模型对损失函数以及训练方式的选择均不相同,而最终均需要输出前景目标的位置、大小以及类别置信度。

目标检测与实例分割任务的性能评估较分类任务更复杂。对于包含N个目标,而检测或实例分割模型输出M个实例的图像,评估阶段通常根据模型输出的检测框或掩膜与ground truth标注的交并比(intersection over union,IoU)判断是否成功检出/分割。N个实际物体中,成功检出/分割的物体称做真正例(true positive, TP),剩余未检出的物体称为假负例(false negative, FN),TP/N则称做该图像中目标的召回率;类似地,在M个输出实例中,正确对应于真实目标的也为TP个真正例,而剩余无效的输出实例称做假正例(false positive, FP),对于模型输出,TP/M称做准确率。

由于分类置信度的存在,筛除部分低置信度的输出实例可能有助于提升准确率。然而,对召回率而言,这种筛除方式可能导致某些正确实例被删去而损害召回性能。因此,随着不同置信度阈值的选取,准确率与召回率之间存在权衡折中。一般地,在目标检测或实例分割评估中,给定某IoU阈值确定检出标准后,便可选取分布于(0,1)内的多个置信度阈值,在2维平面内,绘制以召回率R(recall)为横坐标,准确率P(precision)为纵坐标的散点,并连接成为P-R曲线。通过积分等方式,计算召回率从0~1范围内曲线下方的覆盖面积,便可获得检测与分割任务最为普适的指标之一——平均准确率AP(average precision);而计算不同的IoU阈值下AP的平均值,则能够获得评价较为综合的指标mAP(mean average precision)。通常,不同的数据集或实验配置可能参照不同的评估指标,例如,AP50指标要求固定IoU阈值为0.50以计算AP,而AP50:95指标则要求从0.50开始选取IoU阈值,步进0.05至0.95,以计算此阈值区间的mAP。另一类似的评估指标是平均召回率AR(average recall),由Microsoft COCO(common objects in context)数据集(Lin等,2014)首次提出,指一定输出实例数量下的最大召回率。该指标在检测与分割方案的性能比较中应用相对较少。

语义分割任务中，分割模型$f_{ss}(\cdot)$需要输出与原图$\boldsymbol{I}$同等大小的掩膜$\boldsymbol{M}$。具体为

$$\boldsymbol{M} = f_{ss}(\boldsymbol{I}) \tag{4}$$

掩膜中每个像素对应的向量（按通道维度计）代表其分配至某一类别的概率。由于输入输出像素是一一对应的，因此全卷积的结构被充分应用于不同的语义分割器中。全卷积网络（Long 等，2015）利用卷积与反卷积获得与原图分辨率一致的输出，SegNet（Badrinarayanan 等，2017）利用跳跃连接逐步恢复分辨率，而空洞卷积与条件随机场（conditional random field，cRF）也被引入以获得多尺度信息和像素间的相关性信息（Chen 等，2016a）。

从输入输出角度，语义分割可视做逐像素分类问题，评估指标基于像素进行定义。平均 IoU（mIoU）要求对图像中存在的各类别分别计算真正例 TP、假正例 FP 以及假负例 FN 像素数量，并计算单类别 IoU。具体为

$$IoU = \frac{TP}{TP + FP + FN} \tag{5}$$

计算所有类的 IoU 平均值，可获得平均指标 mIoU。另一类面向全局的评估指标是像素准确率（pixel accuracy，PA），即分类正确的像素数量占总像素数量的比率；同时，若对各类别分别计算 PA 后求平均值，可获得相应的平均指标即平均 PA（mean PA，mPA）。

1.2　分类、检测、分割任务的常用数据集

简单的图像分类任务对灰度图像进行处理，代表性数据集为美国国家标准与技术研究院数据集（mixed national institute of standards and technology database，MNIST）（LeCun 等，2013），收集了大量手写数字并转换为灰度图像，类别为 0～9，共 10 类。数据集共包含 60 000 幅训练图像与 10 000 幅测试图像。自 LeNet（LeCun 等，1998）网络成功将 MNIST 数据集的 Top-1 错误率降至 1.7% 后，各类新颖的深度模型不断出现，现已将该指标降至 0.2% 以内。

利用更通用的图像分类网络对 RGB 彩色自然图像进行处理。ImageNet（Deng 等，2009）数据集是一个大规模通用的自然图像分类数据集，共包含接近 1 500 万幅自然图像与超过 2 万类图像类别，且使用 WordNet 架构进行层次化类别标注。常用于实际任务中的是 ILSVRC（ImageNet Large Scale Visual Recognition Challenge）竞赛从中截取的子集，其训练集包含超 120 万幅图像，验证集与测试集分别包含 5 万与 10 万幅图像，共 1 000 个类别。此数据集已成为对分类模型进行评估的最典型数据集。其余的部分主流分类数据集包括数据量较小的 CIFAR（Canadian Institute for Advanced Research）（Krizhevsky 和 Hinton，2009）系列数据集以及大规模的 WebVision 数据集（Li 等，2017b）等。

面向目标检测与语义分割任务的典型数据集之一为 PASCAL VOC（pattern analysis，statistical modeling and computational learning visual object classes）（Everingham 等，2010）系列，其自 2005—2012 年举行了多次目标检测挑战赛，并对数据集版本进行多次更新，使用最广泛的是 VOC 2007 以及 VOC 2012 数据集。VOC 2012 数据集包含 4 个大类共 20 个前景目标类别，其中 trainval 子集包含 11 530 幅图像，共 27 450 个目标边界框，以及 6 929 张分割掩膜。PASCAL VOC 的体量对比逐渐增长的任务需求，尚属于偏小的集合类型，而大规模检测、分割数据集的典型代表为 Microsoft COCO（Lin 等，2014）数据集，其中除典型的目标检测、实例分割标注外，也包含关键点检测、全景分割等视觉任务的标注。COCO 数据集共含 80 个用于检测与分割任务的有效目标类别。COCO 2017 训练集拥有超过 12 万幅图像与约 150 万个检测与分割实例的标注，其验证集与测试集已成为检测、分割任务的基准评估数据集。更大规模的数据集，例如 Objects365（Shao 等，2019）以及 Open Image 数据集（Krasin 等，2018）等逐渐出现，但均不如 COCO 数据集使用广泛。

上述检测与分割的代表数据集同样包含部分语义分割标注。在 VOC 2012 的语义分割标注中，训练、验证与测试集均包含超 1 400 幅图像，共 21 类；在其图像基础上，SBD（semantic boundaries dataset）（Hariharan 等，2011）数据集对其语义标注进行了扩充。基于来自 VOC 数据集的图像，SBD 数据集增添了大量语义分割标注，包含超 8 000 幅训练图像与超 2 000 幅的测试图像。MS COCO 数据集的语义分割标注来自其物件（stuff）标签，其对 91 类不适用于计数任务的类别进行了分割掩膜标注，并可与 80 类的实例分割标注共同用于全景分割任务。Cityscapes（Cordts 等，2016）为另一个常用的语义分割数据集，主要面向自动驾驶场景，共含 2 975 幅训练图像、500 幅验证图像与 1 525 幅测试图像，包含 19 个类别的语

义分割标注，其中包括含有实例分割标注的8个类别。

1.3 模型蒸馏的基本框架与评价指标

如图1所示，模型蒸馏的基本框架通常由两个预训练网络构成。一个大规模、结构复杂的教师模型$f_{\mathrm{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}})$和一个轻量级、结构简单的学生模型$f_{\mathrm{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}})$。其中，$\boldsymbol{w}_{\mathrm{T}}$与$\boldsymbol{w}_{\mathrm{S}}$分别为教师与学生模型的权重等参数。进行蒸馏时，每次对于相同输入图像$\boldsymbol{I}$，教师与学生模型的（中间或末端）输出分别为$\boldsymbol{y}_{\mathrm{T}}=f_{\mathrm{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}})$以及$\boldsymbol{y}_{\mathrm{S}}=f_{\mathrm{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}})$，二者用于计算特殊设计的蒸馏损失函数，以使二者逼近，具体为

$$L_{\mathrm{dist}}=L(\boldsymbol{y}_{\mathrm{T}},\boldsymbol{y}_{\mathrm{S}}) \tag{6}$$

式中，常见的损失函数$L(\cdot,\cdot)$包括交叉熵损失、KL散度、L2损失函数等。训练过程中，教师模型的参数固定不变，仅在学生模型上进行反向传播得到$\nabla_{\boldsymbol{w}_{\mathrm{S}}}L_{\mathrm{dist}}$，从而与利用真实标签$y$进行常规训练的损失函数$L_{\mathrm{train}}(y,\boldsymbol{y}_{\mathrm{S}})$（例如交叉熵损失）并行，通过总梯度对学生权重$\boldsymbol{w}_{\mathrm{S}}$进行更新，即

$$\boldsymbol{w}_{\mathrm{S}}^{(\mathrm{next})}=\boldsymbol{w}_{\mathrm{S}}-\nabla_{\boldsymbol{w}_{\mathrm{S}}}(L_{\mathrm{dist}}+L_{\mathrm{train}}) \tag{7}$$

式中，$\boldsymbol{w}_{\mathrm{S}}^{(\mathrm{next})}$为更新后的$\boldsymbol{w}_{\mathrm{S}}$，以此便将$f_{\mathrm{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}})$的知识通过$L_{\mathrm{dist}}$迁移至$f_{\mathrm{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}})$，完成知识蒸馏过程。

当前，业界对深度模型的压缩与加速尚无统一评估标准。有关工作如剪枝（Han等，2015；Park等，2017；Tung和Mori，2018）、网络量化（Wu等，2016；Zhou等，2017；Han等，2016）、低秩分解（Sainath等，2013；Kim等，2016）等在评估时往往报告模型压缩前后的参数量、推断或训练速度以及在特定任务上的性能以进行比较。然而，本文介绍的各类模型蒸馏方案，教师模型、学生模型的选取、实验细节相互存在差异。因此，本文尽可能选择公开文献中采用相似模型与数据集进行实验的工作，在给出模型参数量的条件下，利用绝对性能提升进行实验评估。在模型蒸馏中，对教师模型$f_{\mathrm{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}})$而言，使用蒸馏后的轻量级学生模型$f_{\mathrm{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}})$进行替换可达到模型压缩、加速的目的；而对学生模型而言，蒸馏的作用等效于对其自身性能的额外提升。例如，图像分类中对Top-1和Top-5准确率的提升Δacc，目标检测与实例分割中对mAP的提升ΔmAP，语义分割中的ΔmIoU等。考察蒸馏前后的学生模型在上述性能指标上的提升，是评估模型蒸馏性能的典型方法。本文基于此类指标进行模型蒸馏方法的性能评价。

2 分类任务中的模型蒸馏

Hinton等人（2015）最初提出模型蒸馏概念时以分类网络为例进行描述。而后各类模型蒸馏的延伸工作也通常面向分类网络。本节主要介绍典型用于分类网络的模型蒸馏算法，梳理迄今为止的发展历程，并对不同算法的性能进行比较。

基于研究的具体内容，针对分类器模型蒸馏的相关工作大多属于两类。一类对分类器蒸馏的算法进行新的设计，例如设计新的损失函数、用于教师和学生逼近的模型输出等；另一类则将现存蒸馏方法尝试应用于基于图像分类的特定任务。

2.1 蒸馏方式的设计

2.1.1 基础蒸馏方法

本文在引言中介绍，最基础的模型蒸馏方法基于教师与学生模型的输入—输出变量对（pairs）。早期模型蒸馏方案均利用相同输入样本使教师与学生的输出进行逼近。Buciluǎ等人（2006）提出在多个模型间进行知识迁移的可能，并探索复杂模型对数据进行伪标注的作用，同时参考随机标注、朴素贝叶斯等不同获取伪标注的方案，用以训练一个小型神经网络，使之获取伪标签中的知识。这是最早利用预训练模型进行伪标签标注方法的尝试，由于当时技术所限，神经网络仅作为减少传统模型参数量的替代模型。2014年开始，深度学习获得普遍认可，Ba和Caruana（2014）提出使较浅网络模拟深度网络的设想，设计了利用L2损失函数对深层网络输出进行回归的算法，同时提出利用额外线性层以减弱浅层网络与深度网络参数量差异的方案，从而加速模拟算法的收敛。此项研究首次将利用已知模型的输出训练其他模型的思想应用于深度模型，在CIFAR-10（Krizhevsky和Hinton，2009）分类数据集上进行实验并显著提升了部分小型网络的性能。

Hinton等人（2015）引入知识蒸馏的概念，利用一个预训练的教师分类网络$f_{\mathrm{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}})$的输出概率向量$\boldsymbol{p}_{\mathrm{soft}}$作为软标签，对一个轻量学生分类网络$f_{\mathrm{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}})$通过交叉熵损失函数$L_{\mathrm{dist}}$进行训练，具体为

$$\boldsymbol{p}_{\mathrm{soft}}=f_{\mathrm{T}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{T}}) \tag{8}$$

而损失函数为

$$L_{\text{dist}} = -\boldsymbol{p}_{\text{soft}}^{\text{T}} \log \boldsymbol{p}_{\text{S}} \tag{9}$$

式中，$\boldsymbol{p}_{\text{S}}$为学生模型的分类概率向量。对比通用于分类当中的one-hot标签，软标签包含更多的类间结构化信息。例如，对于类别为猫的实例，其更接近狗等相似类别而较不接近山脉等不相似类别。$\boldsymbol{p}_{\text{soft}}$由于包含$f_{\text{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\text{T}})$预测的各类别概率（置信度），可提供不同类别间的隐藏信息，帮助指导学生模型训练。同时，为控制类别标签的软化程度，Hinton等人（2015）提出可利用教师网络分类器的输出向量$\boldsymbol{y}_{\text{T}}$，并通过调节softmax层温度超参数$T$，以控制概率向量$\boldsymbol{p}_{\text{soft}}$，偏离one-hot的程度，为学生提供合适监督。概率向量具体为

$$\boldsymbol{p}_{\text{soft}} = \text{softmax}\left(\frac{\boldsymbol{y}_{\text{T}}}{T}\right) \tag{10}$$

此方案开启了对模型蒸馏的研究，并提供了对分类器通用的软标签调节方法及蒸馏方式。

Hinton等人（2015）的知识蒸馏方法仅基于分类器输出向量$\boldsymbol{y}_{\text{T}}$构造软标签，形式较为简单。其后，出现了各式各样针对分类器的蒸馏算法。Romero等人（2015）将教师网络中间层输出的特征图$\boldsymbol{v}_{\text{T}}$作为线索，利用L2损失函数使学生网络对应的中间层特征$\boldsymbol{v}_{\text{S}}$对其进行模拟，称做线索学习，同时利用ground truth标签以及上述分类软标签对学生模型进行共同监督；同时，由于设计的学生网络FitNet可能具备与教师网络不同（一般更少）的通道数量，因此设计了额外的回归层$r(\boldsymbol{v}_{\text{S}})$用于匹配通道，进行损失函数的计算。具体为

$$L_{\text{HT}} = \frac{1}{2}\boldsymbol{v}_{\text{T}} - r(\boldsymbol{v}_{\text{S}})_2^2 \tag{11}$$

此方案利用网络的分层特性，给予学生网络常规训练中无法获得的标注指导，在CIFAR-10和CIFAR-100等多个分类数据集上获得了显著的性能提升。

此后，模型蒸馏的新颖方案（Buciluǎ等，2006；Ba和Caruana，2014；Hinton等，2015；Romero等，2015）大多以上述研究为基础。基础模型蒸馏算法构造了利用网络输出$\boldsymbol{y}_{\text{T}}$或隐藏层输出$\boldsymbol{v}_{\text{T}}$使学生模型进行逼近的通用思路，方法朴素简单，易于实现，然而优化训练不够灵活，也未考虑逼近任务对知识传递的有效性。

2.1.2　基于互信息的蒸馏

基础模型蒸馏算法基于网络输出与特征的逼近，等效于利用函数的输出进行拟合。而学生与教师模型的输出分布必然存在差异，使不同分布的分类向量或特征进行逼近可能造成优化困难。因此，研究人员对教师—学生的模型输出关系进行重新建模，利用互信息表示教师对学生模型的影响，从而不同于单纯函数拟合，而是基于互信息进行蒸馏。

Ahn等人（2019）利用变分推断对学生至教师的互信息$I(\boldsymbol{v}_{\text{T}};\ \boldsymbol{v}_{\text{S}})$进行最大化，自模型输出特征服从的分布入手，重新建模了模型间监督关系。其构造一个高斯分布，以推断已知学生输出特征的条件下教师输出特征的条件分布$q(\boldsymbol{v}_{\text{T}}|\boldsymbol{v}_{\text{S}})$，通过对该分布的对数求联合期望$E_{\boldsymbol{v}_{\text{T}},\boldsymbol{v}_{\text{S}}}[\log q(\boldsymbol{v}_{\text{T}}|\boldsymbol{v}_{\text{S}})]$的方式最大化互信息的证据下界，从而最大化互信息以达到蒸馏目的。Passalis等人（2020）利用N_L层网络中不同层的特征$\boldsymbol{v}^{N_L}$与训练目标y之间的互信息I，构造网络推断过程中的信息流，从而考察教师各层对学生的作用，在结构差异显著的教师与学生模型间也可迁移知识。该信息流具体为

$$W = \left[I(\boldsymbol{v}^1,y),\cdots,I\left(\boldsymbol{v}^{N_L},y\right)\right]^{\text{T}} \in \mathbf{R}^{N_L} \tag{12}$$

特别地，作者利用基于核函数$\varphi(\boldsymbol{v}_i,\boldsymbol{v}_j)$的二次互信息将知识迁移问题转化为条件概率的匹配问题，再通过Jeffreys散度计算模型特征之间概率匹配的差异程度，以寻找特征层间最合适的监督关系。其中，$\boldsymbol{v}_i$与$\boldsymbol{v}_j$表示不同输入样本的对应层特征。

近年来，对比学习作为一种全新的自监督学习方式获得大量的研究。Tian等人（2022）将对比学习方案与基于互信息的建模结合用于模型蒸馏。计算蒸馏损失时，采样不同的输入样本，对模型输出$f_{\text{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\text{S}})$与$f_{\text{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\text{T}})$的联合分布$p(f_{\text{T}},f_{\text{S}})$建模，通过最大化其在$N$个输入样本下的互信息下界，将类别标注$y=1$的情况视做正样本，以构造对比表征的损失函数进行训练，完成对互信息的最大化。其中互信息下界约束具体为

$$I(f_{\text{T}};f_{\text{S}}) \geqslant \log N + E_{q(f_{\text{T}}f_{\text{S}}|y)}\left[\log q\left(y\middle|f_{\text{T}},f_{\text{S}}\right)\right] \tag{13}$$

式中，q为对f_{T}、f_{S}与标注y三者间各类联合分布或条件分布的通用表示，$E_{q(f_{\text{T}},f_{\text{S}}|y)}[\cdot]$为条件期望。

利用互信息的蒸馏，核心在于对教师、学生模型输出变量间的关系重新建模，并抛弃单纯函数逼近的方式，避免了使分布差异较大的输出变量进行直接逼近而引入的内在不合理性。其不足在于，互信

息中信源与信宿的变量选取以及优化方式的选择缺少参考依据;同时,此类方案在复杂模型与数据集上的通用性尚未验证,蒸馏获得的性能提升相对有限。

2.1.3 基于多教师模型的蒸馏

常规“预训练+蒸馏”方案中,通常仅存在单一的教师与学生模型,除真实标签外,蒸馏算法通常使学生模型利用唯一的教师模型输出进行单次蒸馏训练。对此,研究人员尝试对蒸馏的流程进行适当细化,探索利用多个教师模型进行蒸馏的策略,通过利用多次迭代训练等方法提升蒸馏性能。

Mirzadeh等人(2020)在参数分布、结构差异较大的蒸馏框架中引入了一个作为中介的助教模型$f_{\mathrm{TA}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{TA}})$,其复杂度介于$f_{\mathrm{T}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{T}})$与$f_{\mathrm{S}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{S}})$之间,通过教师蒸馏助教、助教蒸馏学生的流程进行多次迭代式训练。Jin等人(2019)在教师模型训练过程中,基于贪心算法选取若干不同训练阶段的教师模型作为锚点,使学生模型依次以锚点为教师进行渐进式蒸馏,并将其称做路线约束优化(route constrained optimization, RCO)。该算法计算$f_{\mathrm{T}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{T}})$与$f_{\mathrm{S}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{S}})$的KL(Kullback-Leibler)散度$\mathcal{H}_i$,通过设定其不同迭代周期下的相对差异的阈值$\epsilon_h$,利用贪心算法选取处于特定迭代周期的教师模型,以逐步弥合学生模型与预训练教师模型的差距。具体为

$$h_{ij}=\frac{\mathcal{H}_j-\mathcal{H}_i}{\mathcal{H}_i} \tag{14}$$

式中,i,j均表示迭代周期。

Yang等人(2019)则提出了快照蒸馏算法,使学生模型将过去某个迭代下的模型自身$f_{\mathrm{S}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{S}}^{k_i})$作为教师模型,以其输出作为快照,实时指导当前迭代模型$f_{\mathrm{S}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{S}}^{(i-1)})$的参数更新而获得迭代后的模型$f_{\mathrm{S}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{S}}^{i})$,其中,$k_i$表示某个早于第$i$次的迭代。

上述方案均重复利用了网络先前学习的知识,通过多次蒸馏的方式逐级训练学生模型,并省去常规模型蒸馏中的预训练过程,以少量额外计算消耗为代价获得蒸馏性能的有效提升。

Du等人(2020)考察梯度空间,并对蒸馏问题重新建模,计算存在N_{T}个教师模型的条件下,蒸馏损失的梯度$\nabla_{\boldsymbol{w}_{\mathrm{S}}}L_{\mathrm{dist}}^{n}$对参数优化方向$\boldsymbol{d}$的影响。其设计了一组松弛阈值$\xi_n>0$,以通过向量夹角确定蒸馏是否背离合适优化方向,从而将对学生模型的训练重新建模为约束优化问题,具体为

$$\begin{aligned}&\min_{d,g,\xi_n} g+\lambda_0\sum_{n=1}^{N_{\mathrm{T}}}\xi_n+\frac{1}{2}\|\boldsymbol{d}\|^2\\&\text{s.t.}\left\langle\nabla_{\boldsymbol{w}_{\mathrm{S}}}L_{\mathrm{dist}}^{n},\boldsymbol{d}\right\rangle\leqslant g+\xi_n\end{aligned} \tag{15}$$

式中,L_{dist}^{n}表示第n个教师模型的蒸馏损失,g为判别梯度优化方向是否偏离训练的参考量,λ_0为平衡因子。此训练方式面向存在多个教师模型的场景,利用优化问题的重新建模,即时改变教师对学生的监督权重。

ResNet(residual neural network)(He等,2016)系列中,残差模块的提出显著提升了分类网络性能。而Li等人(2020a)尝试将残差块包含的额外知识迁移至常规CNN模块,将仅由常规CNN模块组成的网络作为学生模型,而使包含残差模块且深度相同的ResNet作为教师,采用两个模型相互连接的前向预测方式,使第i个CNN模块输出的特征$f_{\mathrm{CNN}}^{(i)}\left(f_{\mathrm{CNN}}^{(i-1)};\boldsymbol{w}_{\mathrm{S}}^{(i)}\right)$传入对应ResNet模块的下一层,即计算$f_{\mathrm{Res}}^{(i+1)}\left(f_{\mathrm{CNN}}^{(i)};\boldsymbol{w}_{\mathrm{T}}^{(i+1)}\right)$,其中,$i+1$表示第$i$个模块的下一模块。以此基于每次前向预测输出向量$\boldsymbol{y}_{\mathrm{T}}$与概率软标签$\boldsymbol{p}_{\mathrm{soft}}$进行蒸馏,便可将多个$f_{\mathrm{Res}}^{(i+1)}(\cdot;\cdot)$模块的知识传递至对应的$f_{\mathrm{CNN}}^{(i+1)}(\cdot;\cdot)$。

基于多教师模型的蒸馏与常规的“教师输出+伪标签训练”模式不同,该方案是利用更细致的训练流程,将来自不同教师模型的信息分步传递至学生。此类方案能够自适应地根据蒸馏阶段不同,筛选教师输出以指导学生训练,从而更充分地利用预训练模型知识。然而,此类方案通常导致模型蒸馏复杂度显著提升。

2.1.4 基于后处理的蒸馏

网络隐藏层的特征图$\boldsymbol{v}_{\mathrm{T}}$常用于模型蒸馏中的线索学习,其不同位置、通道的像素包含重要程度不同的信息,而教师模型的输出的分类向量$\boldsymbol{p}_{\mathrm{soft}}$也非完美标签,因此,对$\boldsymbol{p}_{\mathrm{soft}}$与$\boldsymbol{v}_{\mathrm{T}}$进行后处理可能获取更合适的蒸馏监督信号,有助于提升学生模型泛化性。因此,研究人员尝试对特征图与软标签进行不同后处理,以用于提出的模型蒸馏算法。

Sau和Balasubramanian(2016)提出将教师的输出软标签$\boldsymbol{p}_{\mathrm{soft}}$添加高斯噪声扰动$\boldsymbol{\xi}$后,再作为损失函数的额外正则化项,利用L2损失训练学生模型,即

$$L_{\mathrm{dist}}=\left\|\frac{1}{2}f_{\mathrm{S}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{S}})-(1+\boldsymbol{\xi})\boldsymbol{p}_{\mathrm{soft}}\right\|_2^2 \tag{16}$$

式中，经后处理的$(1+\xi)\boldsymbol{p}_{\text{soft}}$包含随机变量成分，构成了带有正则化作用的监督信号，用以引导学生模型的分类输出$f_{\text{S}}(\boldsymbol{I};\boldsymbol{w}_{\text{S}})$。Sau 和 Balasubramanian（2016）对比了此方案与直接对学生网络参数使用 L2 正则化项的训练方式，并说明相较于直接在学生模型参数添加噪声扰动，通过包含噪声的软标签进行蒸馏能够获得更优的性能。

Zagoruyko 和 Komodakis（2017a）对模型的特征图进行后处理。其考察模型蒸馏方案对单张特征图不同位置的关注程度，引入了视觉模型中考察区域重要性的注意力机制，通过多通道归一化的映射$F:\mathbf{R}^{C\times H\times W}\rightarrow\mathbf{R}^{H\times W}$计算特征图的多个通道的总特征，并依此分别计算了基于激活层输出的注意力图与基于梯度的注意力图，用以监督学生模型相应的此两类注意力图。基于激活层的注意力图中，利用维度为$\mathbf{R}^{C\times H\times W}$的特征图计算维度为$\mathbf{R}^{H\times W}$的注意力图后，再利用$p$范数损失使归一化的二者逼近；另一方面，考察教师模型当前参数为$\boldsymbol{w}_{\text{T}}$的条件下，损失函数$L$对各个特征像素$v$上的偏导数$\frac{\partial L}{\partial v}$，以此获得梯度注意力图$\boldsymbol{M}_{\text{T}}$，再利用 2 范数损失使学生模型生成的$\boldsymbol{M}_{\text{S}}$向教师的$\boldsymbol{M}_{\text{T}}$逼近。通过上述方案，教师与学生特征图$\boldsymbol{v}_{\text{S}}$与$\boldsymbol{v}_{\text{T}}$经后处理输出的注意力图完全取代了$\boldsymbol{v}_{\text{S}}$与$\boldsymbol{v}_{\text{T}}$本身，用于进行线索学习。

基于后处理的蒸馏算法对模型输出的特征图$\boldsymbol{v}_{\text{T}}$或软标签$\boldsymbol{p}_{\text{soft}}$进行后处理，再作为蒸馏监督信号，选择性地使学生学习教师输出中包含的有效信息。其优势在于避免了学生对特征或软标签的单调模拟，以正则化作用缓解了共同优化带来的优化分歧问题；而其劣势除复杂度增大以外，还包含后处理方法难以量化的问题。例如，软标签扰动噪声的方差大小、构造注意力图选取的维度及变量等均无法在训练前预估其对蒸馏性能的影响。

2.1.5　基于样本间关系的蒸馏

模型隐藏层特征$\boldsymbol{v}_{\text{T}}$与末端的分类向量$\boldsymbol{p}_{\text{soft}}$均属于网络的局部或末尾输出；上一小节中，用于蒸馏的模型输出虽经过后处理，但是其仍然对应单一样本（输入图像、隐藏层）。而考察不同输入图像、网络不同层的特征之间的关系，能够构造更合适的表征，以迁移来自教师的知识。

考虑到深度神经网络不同层之间特征的关系，Yim 等人（2017）提出基于同一网络内不同层之间的特征图联合表示网络所具备的知识。首先，对教师$f_{\text{T}}(\boldsymbol{I};\boldsymbol{w}_{\text{T}})$与学生$f_{\text{S}}(\boldsymbol{I};\boldsymbol{w}_{\text{S}})$均分别取特定层输出的若干组特征，每组包含两张特征图$\boldsymbol{v}^1\in\mathbf{R}^{C_1\times H\times W}$以及$\boldsymbol{v}^2\in\mathbf{R}^{C_2\times H\times W}$，将其分别按逐通道像素在$H\times W$维度展平为 1 维向量后，仿照计算向量组格拉姆矩阵的方式，获得解决过程流（flow of the solution procedure，FSP）矩阵$\boldsymbol{A}\in\mathbf{R}^{C_1\times C_2}$，以此表示每个模型的知识。具体为

$$\boldsymbol{A}_{i,j}(\boldsymbol{I};\boldsymbol{w}_{\text{S}})=\sum_{a=1}^{H}\sum_{b=1}^{W}\frac{\boldsymbol{v}^1_{i,a,b}(\boldsymbol{I};\boldsymbol{w}_{\text{S}})\times\boldsymbol{v}^2_{j,a,b}(\boldsymbol{I};\boldsymbol{w}_{\text{S}})}{H\times W}\tag{17}$$

式中，i,j表示由两幅特征图中不同通道的采样计算获得的矩阵中第i行第j列元素的值。然后，使学生模型的每组 FSP 矩阵$\boldsymbol{A}(\boldsymbol{I};\boldsymbol{w}_{\text{S}})$通过加权 L2 损失对教师相应组的$\boldsymbol{A}(\boldsymbol{I};\boldsymbol{w}_{\text{T}})$进行拟合。最后，通过在 CIFAR 系列数据集上的实验，验证了其方案加速优化（收敛）及提升学生模型性能的能力。

除对不同层特征图之间的关系进行考察，研究人员对不同输入图像对应的模型输出特征进行研究。Park 等人（2019）利用三元组集合方式，对训练集中不同图像进行采样，分组组合为三元组特征$[\boldsymbol{v}_i,\boldsymbol{v}_j,\boldsymbol{v}_k]$，方括号中 3 个变量代表采样的任意 3 个样本（图像）对应的模型输出特征，三者共同构成单一三元组。通过考察教师、学生模型对应相同三元组的特征，分别计算特征间归一化欧氏距离和三元组特征两两差值的余弦相似度。归一化欧氏距离具体为

$$\psi_D(\boldsymbol{v}_i,\boldsymbol{v}_j)=\frac{1}{\mu}\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2\tag{18}$$

余弦相似度具体为

$$\psi_A(\boldsymbol{v}_i,\boldsymbol{v}_j,\boldsymbol{v}_k)=\left\langle\frac{\boldsymbol{v}_i-\boldsymbol{v}_j}{\left\|\boldsymbol{v}_i-\boldsymbol{v}_j\right\|_2},\frac{\boldsymbol{v}_j-\boldsymbol{v}_k}{\left\|\boldsymbol{v}_j-\boldsymbol{v}_k\right\|_2}\right\rangle\tag{19}$$

式中，归一化因子μ为相异样本对的数量。然后，分别利用 L2 与 L1 损失函数使$f_{\text{S}}(\boldsymbol{I};\boldsymbol{w}_{\text{S}})$中构造的三元组在$\psi_D$与$\psi_A$两种表征上逼近教师$f_{\text{T}}(\boldsymbol{I};\boldsymbol{w}_{\text{T}})$的对应表征。

Tung 和 Mori（2019）则利用具有批规模b的单个输入批次中不同图像的特征，进行矩阵乘积与归一化操作，获得与式（17）中 FSP 类似（此处针对不同图像，FSP 针对不同层）的相似度矩阵$\boldsymbol{A}=Normalize(\boldsymbol{v}\boldsymbol{v}^{\text{T}})$，其中，$\boldsymbol{v}$为同一批次的输出特征；然后，利用 L2 损失函数使学生与教师相似度矩阵逼

近。Peng等人(2019)的策略与上述基于相似度的方案(Tung和Mori,2019)十分类似,额外引入相关性度量以代替简单的矩阵乘积。具体为

$$\boldsymbol{C}_{ij}=\varphi\left(\boldsymbol{v}_i,\boldsymbol{v}_j\right) \tag{20}$$

式中,i,j表示任意两个采样。此外,介绍了若干适用于进行该度量的核函数$\varphi(\cdot,\cdot)$。

相比上述工作,Liu等人(2019a)同时考察了不同层与不同样本之间的特征关系。对N幅输入图像,利用网络某层l的特征图$\boldsymbol{v}^l(\boldsymbol{I};\boldsymbol{w})$构建了一个无向完全图,称为实例关系图(instance relation graph,IRG),具体为

$$\boldsymbol{IRG}_l=\left(\left\{\boldsymbol{v}^l\left(\boldsymbol{I}_i;\boldsymbol{w}\right)\right\}_{i=1}^{N},\boldsymbol{E}_l\right) \tag{21}$$

图中各节点对应每幅输入图像的相应特征,图的边权值$\boldsymbol{E}_l$则为欧氏距离平方,即

$$\boldsymbol{E}_l(i,j)=\left\|\boldsymbol{v}^l\left(\boldsymbol{I}_i;\boldsymbol{w}\right)-\boldsymbol{v}^l\left(\boldsymbol{I}_j;\boldsymbol{w}\right)\right\|_2^2 \tag{22}$$

式中,$\boldsymbol{I}_i$与$\boldsymbol{I}_j$表示任意两幅不同图像。对网络不同层,可构建不同的IRG;而对任意不同两层l_1与l_2的IRG,提出一种IRG变换$Trans\left(\boldsymbol{IRG}_{l_1},\boldsymbol{IRG}_{l_2}\right)$。对相同输入样本,通过计算二者的节点特征距离$\left\|\boldsymbol{v}^{l_1}(\boldsymbol{I}_i;\boldsymbol{w})-\boldsymbol{v}^{l_2}(\boldsymbol{I}_i;\boldsymbol{w})\right\|_2^2$与二者邻接矩阵的距离$\left\|\boldsymbol{E}_{l_1}-\boldsymbol{E}_{l_2}\right\|_2^2$,衡量两层特征的差异性。进行蒸馏时,首先考察教师与学生模型单层节点特征之间与边权值之间的L2损失,再考察二者在网络中相异两层通过上述IRG变换获得的两种距离间的差异。由此,在考察模型间、样本间关系的同时,考虑了网络不同层之间特征表示的关系。

基于样本间关系的蒸馏改变了选取单层或单个样本输出进行模拟的方案,转而利用不同层、不同样本或不同位置特征间的相互关系作为监督信号。其利用更高阶的表示,能够传递层间关系等隐式模型知识,但对样本数量以及合适样本的选择有较高要求,更多样本间关系的构造方式有待探索。

2.1.6 基于辅助子网的蒸馏

前文介绍的方案中,训练损失函数均由教师、学生模型自身的输出进行构造。随着对深度神经网络研究的深入,部分其他领域的模型与方法可应用于知识蒸馏的改进,本小节介绍两种利用子网络模块辅助模型蒸馏的方案。

Xu等人(2018)参照条件生成对抗网络(conditional generative adversarial network,cGANs)(Mirza和Osindero,2014)的结构添加辅助子网络。除利用常规模型蒸馏方法使学生输出特征$\boldsymbol{v}_{\mathrm{S}}$逼近$\boldsymbol{v}_{\mathrm{T}}$外,在学生模型末端添加了额外的判决器子网$D(\boldsymbol{v})$,基于GAN的训练思路,使判决器将教师的输出特征$\boldsymbol{v}_{\mathrm{T}}$判为真,而将$\boldsymbol{v}_{\mathrm{S}}$判为假,以此训练学生模型混淆$D(\boldsymbol{v})$的能力,从而使学生模型输出与教师更为接近的特征。Zhang等人(2017)引入知识映射网络(knowledge projection network,KPN)的概念,添加用于映射教师、学生特征间监督关系的子网KPN,将模型蒸馏应用于仅有部分标签的数据集的训练中。利用KPN的输出$f_{\mathrm{KP}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{KP}})$决定教师不同层与学生不同层的对应关系,使知识能够从正确的路径对学生进行监督。因此,在蒸馏过程中同步引入了类似自适应剪枝—微调的操作,利用蒸馏损失函数L_{KP}控制$f_{\mathrm{KP}}(\boldsymbol{I};\boldsymbol{w}_{\mathrm{KP}})$部分的剪枝,在每一轮迭代中,从多个候选KPN内迭代训练获得最合适的一支,用于下一步的蒸馏训练。

判决器子网与KPN的添加,扩展了模型蒸馏方案中仅利用教师、学生分类器自身模型进行输出逼近的思路,利用子网的输出构造损失函数以监督学生模型训练。此类特殊的知识迁移方式对学生自身输出约束较少,却隐式地确保其最终性能逼近教师。此类方法的代价是子网设计较为困难,更难保证训练收敛,同时,基于子网输出的优化任务大幅增加了模型蒸馏复杂度。

2.2 特定任务中的模型蒸馏应用

2.2.1 蒸馏应用:基于分类的进阶视觉任务

不同于2.1节介绍的相关研究,许多关于模型蒸馏的工作未对分类器的蒸馏方案进行新的设计,而是将知识迁移等思想应用于以分类网络为基础或辅助的特定视觉任务,并达到模型压缩或辅助任务进行的目的。

Gupta等人(2016)利用蒸馏方法实现多模态学习中监督信号的迁移,依照与任务蒸馏(Girdhar等,2019)相似的思想,利用在ImageNet(Deng等,2009)数据集上预训练的分类网络中间层获取表示特征$\boldsymbol{v}_{m_1}\left(\boldsymbol{I}_{m_1}\right)$,对利用深度与光流图像等缺少标签的数据对应的输出特征$\boldsymbol{v}_{m_2}\left(\boldsymbol{I}_{m_2}\right)$提供引导,其中,$m_1$与$m_2$分别代表两种模态。该方法中,教师、学生的不同之处在于输入模态,而非模型差异。为对齐不同模态下

的输出特征维度，利用了额外的仿射函数$r\left(\boldsymbol{v}_{m_2}\left(\boldsymbol{I}_{m_2}\right)\right)$改变通道数。

在域迁移领域，Ruder等人（2017）利用模型蒸馏辅助域适应任务，并将一一对应的“教师—学生对”模式推广至使用多个不同数据域上训练的教师$f_{\mathrm{T}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}}^{i}\right)$进行蒸馏的场景（上标$i$表示第$i$个数据域）。利用学生所在数据域$\boldsymbol{D}_{\mathrm{S}}$与各个教师所在数据域$\boldsymbol{D}_{T}$的归一化相似度$sim\left(\boldsymbol{D}_{\mathrm{S}},\boldsymbol{D}_{\mathrm{T}}\right)$，对教师模型的输出软标签$\boldsymbol{p}_{\mathrm{soft}}$进行加权，再用于对学生的蒸馏；而在单一教师的情况下，使用超参数λ以控制教师分类结果引导的硬标签$\boldsymbol{y}_{\mathrm{teacher}}$与教师输出的软标签$\boldsymbol{p}_{\mathrm{soft}}$的叠加权重，从而控制蒸馏标签$\boldsymbol{y}_{\mathrm{dist}}$的软化程度。即

$$\boldsymbol{y}_{\mathrm{dist}}=(1-\lambda)\boldsymbol{y}_{\mathrm{teacher}}+\lambda\boldsymbol{p}_{\mathrm{soft}} \tag{23}$$

Luo等人（2016）在人脸识别问题中应用基础模型蒸馏策略，采样教师与学生模型隐藏层的部分输出特征，并利用L2损失函数使二者逼近；其中，将采样问题视做一个完全图上的推断问题，并最小化由不同神经元$\left\{v_i\right\}_{i=1}^{N_n}$计算的能量函数，从而选择区分能力较强但相互关联较少的神经元。具体为

$$E_{\mathrm{eng}}(\boldsymbol{v})=\sum_{i=1}^{N_n}\Phi\left(v_i\right)+\lambda\sum_{i=1}^{N_n}\sum_{j=1,j\neq i}^{N_n}\Psi\left(v_i,v_j\right) \tag{24}$$

式中，N_n为神经元数量，λ为权重超参数，$\Phi(\cdot)$与$\Psi(\cdot,\cdot)$分别表示一阶与二阶能量函数。

类似地，在行人重识别等需要考察多个样本间关系的任务中，Chen等人（2017b）将特征相似度应用于预测有序样本列表的任务中，利用“教师—学生模型对”预测列表排序的后验概率$\boldsymbol{P}(\pi\in\boldsymbol{P}\mid\boldsymbol{v})$，分别利用由KL散度计算的软损失函数$L_{\mathrm{soft}}$以及由最大似然损失组成的硬损失函数$L_{\mathrm{hard}}$对学生模型进行联合的监督训练，具体为

$$L_{\mathrm{soft}}=\sum_{\pi\in\mathrm{P}}\boldsymbol{P}(\pi|\boldsymbol{v}_T)\log\frac{\boldsymbol{P}(\pi|\boldsymbol{v}_T)}{\boldsymbol{P}(\pi|\boldsymbol{v}_S)} \tag{25}$$

以及

$$L_{\mathrm{hard}}=-\log\boldsymbol{P}(\pi_y|\boldsymbol{v}_T) \tag{26}$$

式中，$\boldsymbol{P}$为所有可能的排序集合，π代表$\boldsymbol{P}$中不同排序的采样。

在基于分类的其他视觉任务或部分需要进行知识迁移的任务中，模型蒸馏的引入额外提升了此类任务的性能，效率与性能的折中较为平衡，难点在于设计从分类任务到应用蒸馏的具体任务的迁移方式，以及确定不同任务特征的对应关系。

2.2.2　结合模型压缩方法的蒸馏应用

作为模型压缩的方法之一，模型蒸馏也可与其他压缩方案结合。Wang等人（2016）提出了支配卷积核（dominant kernel，DK）的概念以进行卷积核的低秩分解，获得运算速度更快的学生模型$f_{\mathrm{LR}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{LR}}\right)$。同时，运用$f_{\mathrm{T}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}}\right)$输出的分类软标签$\boldsymbol{p}_{\mathrm{soft}}$与隐藏层特征$\boldsymbol{v}_{\mathrm{T}}$蒸馏学生网络，以弥补网络轻量化的性能损失。类似地，Li等人（2020b）首先通过剪枝方法剪除教师模型$f_{\mathrm{T}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}}\right)$的部分通道以构造轻量学生模型$f_{\mathrm{pruned}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{pruned}}\right)$，然后利用额外卷积层$Conv(\cdot)$对齐二者的通道数后，再将对齐后的学生特征$Conv\left(\boldsymbol{v}_{\mathrm{pruned}}\right)$进行蒸馏，最后将$Conv(\cdot)$与其前置位的卷积操作通过线性乘积合并为同一层，用于轻量学生模型的推断。Polino等人（2018）与Mishra和Marr（2017）均将网络量化与模型蒸馏进行结合，利用全精度的教师模型$f_{\mathrm{T}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}}\right)$指导量化后的低精度学生模型$f_{\mathrm{Q}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{Q}}\right)$训练。其中，Polino等人（2018）采用迭代式方法进行：对学生网络进行“量化—蒸馏—量化”的重复过程，而教师模型保持全精度权重不变，与基于微调的量化方案类似；Mishra和Marr（2017）则提出与常规蒸馏不同的方案：在蒸馏过程中，教师$f_{\mathrm{T}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}}\right)$本身亦同步进行训练，因此在自身性能的逐渐提升中，$f_{\mathrm{T}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}}\right)$能够将知识传递至下游的低精度学生模型。Crowley等人（2018）则将轻量级网络设计方案与模型蒸馏进行结合，简化$f_{\mathrm{T}}\left(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}}\right)$的卷积操作以获得学生模型，同时利用与Polino等人（2018）相似的训练方式进行迭代式的“压缩—蒸馏”循环步骤；对于蒸馏损失函数的构造，除利用软标签$\boldsymbol{p}_{\mathrm{soft}}$进行监督外，作者亦使用类似Zagoruyko和Komodakis（2017a）的注意力迁移机制构造激活层的注意力图$M_{\mathrm{att}}\in\mathbf{R}^{H\times W}$辅助蒸馏过程，最终提升轻量学生模型的整体性能。

模型蒸馏具有良好的性能提升能力，因此其能够作为其他模型压缩方案的末尾步骤，以弥补量化等方法对视觉模型性能的大幅削弱，从而获得效率、性能均足够良好的压缩模型。其应用中的缺点则在于，对压缩过后有较大结构改变或参数精度改变的模型，蒸馏的微调较为困难，且一般会引入额外复杂度。

2.2.3　改良模型训练方式的蒸馏应用

部分工作中，研究人员不满足于利用ground truth直接进行损失函数计算的常规模型训练方式，因此模型蒸馏亦用于对模型训练方式的改良。如引言中所述，模型蒸馏属于特殊的迁移学习，因此Chen等人（2016b）考察利用预训练模型的方式时，在基础的初始化部分令新生成的学生网络$f_{\mathrm{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}})=f_{\mathrm{T}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{T}})$，以此对学生网络进行权重初始化，加速后续任务的收敛。但该方案更多侧重于权重矩阵$\boldsymbol{W}_{\mathrm{T}}$到$\boldsymbol{W}_{\mathrm{S}}$的转化，而非利用教师模型输出辅助训练。在对无标签数据进行“标注—二次训练”的半监督任务中，Wang等人（2020a）采用了Mixup（Zhang等，2018）算法中提出的数据增强方法，求解优化问题

$$\min_{\lambda}\max_{y} p\left(\boldsymbol{p}_{\mathrm{S}}|\boldsymbol{I}\right),\ \text{s.t.}\ \lambda\in[0,1] \tag{27}$$

以扩充合适的增强样本，从一个黑盒模型中获得无标签图像的软标签$\boldsymbol{p}_{\mathrm{soft}}$，进行学生网络的二次训练。其中$p\left(\boldsymbol{p}_{\mathrm{S}}|\boldsymbol{I}\right)$为已知输入图像$\boldsymbol{I}$时，输出类别向量$\boldsymbol{p}_{\mathrm{S}}$的条件概率，$y$为类别标签，$\lambda$则为Mixup数据增强的权重超参数。在该过程中，发现使用$\boldsymbol{p}_{\mathrm{soft}}$作为标注的训练性能甚至超过了部分人类手工标注的训练性能。面向标注完整的场景，Yun等人（2020）利用相同类别标签y的不同输入样本$\boldsymbol{I}$与$\boldsymbol{I}'$分别输入学生与教师模型，获得$\boldsymbol{p}_{\mathrm{S}}=f_{\mathrm{S}}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}})$及$\boldsymbol{p}_{\mathrm{soft}}=f_{\mathrm{T}}(\boldsymbol{I}';\ \boldsymbol{w}_{\mathrm{T}})$，再计算二者的KL散度以进行训练；由于软标签来自与学生网络的输入$\boldsymbol{I}$不同的图像$\boldsymbol{I}'$，因此这种训练方式提升了学生模型的泛化能力。

上述工作中，模型蒸馏作为额外任务参与非常规的模型训练，进行了任务所需的知识迁移。然而，此类工作对训练方式的改良，通常进行了特殊设计，缺少共性，因此利用蒸馏对训练方式的改善方法难以推广。

2.2.4　分布式训练与加速场景的蒸馏应用

对网络训练的硬件平台考察是模型蒸馏的应用场景之一。Anil等人（2020）在多GPU上进行神经网络分布式训练时，利用了模型蒸馏方法，对一个已经完成并行初始训练（burn in）阶段的网络$f_{\mathrm{S}_0}(\boldsymbol{I};\ \boldsymbol{w}_{\mathrm{S}_0})$，在多个GPU上对自身进行并行的“共蒸馏”（codistillation）。训练过程中，当任意GPU的模型迭代时，选取其余GPU模型输出的平均特征对当前GPU的模型进行监督；由于初始的训练阶段中，各GPU的模型$f_{\mathrm{S}_i}(\cdot;\ \boldsymbol{w}_{\mathrm{S}_i})$均输入不同数据切片$\boldsymbol{I}_i$，因此训练获得的模型参数也不相同；蒸馏过程能够并行进行，保证训练速度，同时能够充分利用蒸馏损失L_{dist}聚合不同GPU之间的信息，即

$$L_{\mathrm{dist}}=L_{\mathrm{hard}}\left(f_{\mathrm{S}_i},\boldsymbol{y}\right)+L_{\mathrm{soft}}\left(f_{\mathrm{S}_i},\frac{1}{N_{\mathrm{GPU}}-1}\sum_{j\neq i}f_{\mathrm{S}_j}\right) \tag{28}$$

式中，L_{hard}与L_{soft}分别为当前GPU的常规训练损失与利用其余GPU输出计算的蒸馏损失，N_{GPU}为GPU数量。基于上述分布式训练方式，网络权重的优化最终得以收敛。在对集成学习的研究中，Malinin等人（2019）以贝叶斯视角对集成特征的概率分布$p(\boldsymbol{v}|\boldsymbol{I})$建模，引入输出分布的隐分布采样$\boldsymbol{\pi}$，使用互信息$\boldsymbol{I}(\boldsymbol{v},\boldsymbol{\pi}|\boldsymbol{I};\ \boldsymbol{w})$优化模型参数在数据集$\boldsymbol{D}$下的条件分布$p(\boldsymbol{w}|\boldsymbol{D})$，使模型在蒸馏中学习多个集成的分布而非其绝对数值。

除分布式训练外，在卷积网络浅层接入分类器以减少推断时间是有效的加速方案。此类方案改变了传统分类网络中仅在主干网络末尾接入全连接分类器的设置，在卷积网络的多个中间层接入了N_{cls}个分类输出分支$\left\{Class\left(\boldsymbol{v}^i\right)\right\}_{i=1}^{N_{\mathrm{cls}}}$，使模型推断时可参考多分支的平均预测结果，或选择其中最合适的预测作为输出。Zhang等人（2019）训练此类多层输出网络时，同步使用最深卷积层输出的特征$\boldsymbol{v}$与软标签$\boldsymbol{p}_{\mathrm{soft}}$对模型自身的浅层输出$\boldsymbol{v}^i$与$Class\left(\boldsymbol{v}^i\right)$进行蒸馏，使浅层网络保有推断速度优势的同时，学习深层特征更丰富的语义信息。由于该方案中教师模型为学生自身，因此将此方案称做自蒸馏。Phuong和Lampert（2019）将基于此类蒸馏方法的网络称为多出口架构，更直接地利用不同的深层卷积层输出的特征$\boldsymbol{v}_j$，对浅层特征$\boldsymbol{v}_i$进行基于交叉熵损失CE(softmax($\boldsymbol{v}_i$), softmax($\boldsymbol{v}_j$))的蒸馏，以充分传递深层卷积的特征提取能力。

在分布式或加速训练的场景下，模型蒸馏事实上仍作为一种性能补偿方法出现，而此类场景通常存在多个教师或学生模型，此类多对多的蒸馏策略提供了对高效、高速模型的性能补偿，难点在于设计多个模型、多层输出的匹配与监督方式。

2.3　分类器蒸馏方法的性能比较

上述各类模型蒸馏方法大多基于分类网络进行设计，验证实验在各类通用的分类数据集上进行。由于不同工作的实验设置存在差异，选取的基线模

型、迭代长度均有所不同，同时其中部分工作是基于分类模型的更复杂的任务，因此，本文在进行实验结果分析时，首先保证数据集的一致性，选取各类模型蒸馏方法在具备良好代表性的ImageNet（Deng等，2009）、CIFAR系列（Krizhevsky和Hinton，2009）分类数据集上的实验性能；其次，对于教师与学生模型，本文均选取在结构相同或近似的网络上进行的实验性能，并报告具体的教师与学生模型类型，从而参照其具体结构，更公平地分析不同方法的性能。在性能指标方面，本文选取实验设置相似的工作进行对比，统一采用分类任务上的绝对性能（以Top-1准确度表示）以及蒸馏后学生模型相对于基线模型的性能提升Δacc进行评估，同时提供模型的参数量信息。

表2—表4列出了不同方案进行模型蒸馏后，学生模型在ImageNet（ILSVRC2012）（Deng等，2009）、CIFAR-10和CIFAR-100（Krizhevsky和Hinton，2009）数据集上获得的Top-1准确率以及蒸馏算法为其带来的提升值Δacc。Δacc指相对于蒸馏前的基线学生模型（如果有）的准确率提升值。分类性能均保留两位小数。教师模型及学生模型栏的类型中，R，X，WR和C分别表示ResNet（He等，2016）、ResNeXt（Xie等，2017）、WideResNet（Zagoruyko和Komodakis，2017b）和不含残差连接的常规卷积网络，MobileV2指MobileNetV2（Sandler等，2019），其他部分简写代表基于特定的模型压缩方法模型。参数量方面，针对不同数据集，同类的模型可能存在规模差异，例如ResNet（He等，2016）系列应用于ImageNet（Deng等，2009）及CIFAR（Krizhevsky和Hinton，2009）系列的模型具有不同的网络结构；而分类类别数量不同会引入全连接层的参数量差异。另外，部分蒸馏方案自行修改了网络的某些结构，可能造成参数量与常规模型所报告的不一致。因此，对于自行报告模型参数量的方案，本文采用其报告的数值，对于未报告的，选取分类模型官方报告的数值或依据其描述的结构计算参数量。为在不同数量级参数量之间规范表示，本文对所有的参数量均保留两位有效数字。

表2　在ImageNet数据集上，面向分类器的部分模型蒸馏方法的实验性能

Table 2　Experiment performances of several model distillation methods for classifiers on ImageNet dataset

蒸馏方法	教师模型			学生模型			Δacc/%
	类型	参数量/M	Top-1准确率/%	类型	参数量/M	Top-1准确率/%	
TA（Mirzadeh等，2020）	R50	25	–	R14	5.6	67.36	2.16
RCO（Jin等，2019）	R50	25	75.49	MobileV2	3.4	68.21	4.01
ResDist（Li等，2020a）	R50	25	76.11	C50	22	76.08	**8.63**
AE-KD（Du等，2020）	R50	25	77.52	R18	11	69.14	0.96
Self-Dist（Zhang等，2019）	R50	25	73.56	R50	25	74.73	1.17
IRG（Liu等，2019a）	R101	44	78.05	X26_32x4d	14	77.18	1.02
CRD（Tian等，2022）	R34	22	73.31	R18	11	71.38	1.68
CS-KD（Yun等，2020）	X101_32x4d	42	**78.40**	X101_32x4d	42	**78.80**	0.40

注：加粗字体表示各列最优结果，“–”表示未报告教师模型性能。

从表2—表4可观察到，基于分类网络的的模型蒸馏均使学生模型的性能取得了可观提升，部分进行自蒸馏的方案相较于模型自身常规训练的基线精度更高。其中，绝对性能最高的方法往往具有较大的模型体量或具有性能极高的教师模型。例如，CS-KD方法（Yun等，2020）利用ResNeXt（Xie等，2017）进行自蒸馏，Zhang等人（2019）则选择ResNet-152作为教师模型。由于部分经过权重量化或经轻量设计的学生网络本身即为性能较低的基线，因此其经过蒸馏通常可获得较高的性能提升。例如，RCO方法（Jin等，2019）对轻量网络进行蒸馏，DistQuant（Polino等，2018）中学生模型为4位权重的网络，经过蒸馏其性能均获得极大提升。

整体上，基于分类器的模型蒸馏中，无论对蒸馏方式进行设计还是在特定任务中进行应用，对教师模型软标签或特定隐藏层特征的获取均较为便利，

表3　在CIFAR-10数据集上，面向分类器的部分模型蒸馏方法的实验性能

Table 3　Experiment performances of several model distillation methods for classifiers on CIFAR-10 dataset

蒸馏方法	教师模型			学生模型			Δacc/%
	类型	参数量/M	Top-1准确率/%	类型	参数量/M	Top-1准确率/%	
VID(Ahn等，2019)	WR40-2	2.2	94.26	WR16-1	0.24	91.85	1.13
cGANs-KD(Xu等，2018)	WR40-10	56	**95.81**	WR10-4	1.2	93.91	1.37
SP(Tung和Mori，2019)	WR16-8	11	95.76	WR40-2	2.2	**95.53**	0.71
DistQuant(Polino 等，2018)	WR28-20	150	95.74	4bits WR28-20	150	94.73	**13.64**
TA(Mirzadeh等，2020)	R110	1.7	–	R8	0.080	88.98	0.46
ResDist(Li等，2020a)	R50	23	95.31	C50	20	94.40	–
A Gift(Yim等，2017)	R26	0.37	91.91	R8	0.080	88.70	0.79
IRG(Liu等，2019a)	R20	1.1*	91.45	R20h	0.28	90.69	1.34
AE-KD(Du等，2020)	R56	0.85	95.78	R20	0.27	93.01	0.51
KP(Zhang等，2017)	R38	3.1	91.15	R50h	0.27	92.37	4.84
FSKD(Li等，2020b)	VGG16	15*	92.66	VGG16-56%	5.3	92.54	7.12
ENSM(Malinin等，2019)	VGG16	140	84.60	VGG16	140	86.80	2.20

注：加粗字体表示各列最优结果，“–”表示未报告教师模型性能或不存在基线学生模型，*标示两处参数量为工作本身所报告，R20及VGG16与官方参数量存在差异可能由于实现结构的不同。

表4　在CIFAR-100数据集上，面向分类器的部分模型蒸馏方法的实验性能

Table 4　Experiment performances of several model distillation methods for classifiers on CIFAR-100 dataset

蒸馏方法	教师模型			学生模型			Δacc/%
	类型	参数量/M	Top-1准确率/%	类型	参数量/M	Top-1准确率/%	
VID(Ahn等，2019)	WR40-2	2.2	74.16	WR40-2	2.2	76.11	1.77
CRD(Tian等，2022)	WR40-2	2.2	75.61	WR16-2	0.79	75.64	2.38
cGANs-KD(Xu等，2017)	WR40-10	56	79.38	WR10-4	1.2	74.25	2.77
DistQuant(Polino等，2018)	WR28-10	37	77.21	4bits WR28-10	37	76.31	2.84
RCO(Jin等，2019)	R50	25	79.34	MobileV2	3.4	70.85	**8.97**
ResDist(Li等，2020a)	R50	25	78.39	C50	22	78.16	–
RKD-DA(Park等，2019)	R50	25	77.76	VGG11	130	74.66	3.40
A Gift(Yim等，2017)	R32	0.46	64.06	R14	0.18	63.33	4.68
TA(Mirzadeh等，2020)	R110	1.7	–	R8	0.080	61.82	0.45
CCKD(Peng等，2019)	R110	1.7	–	R20	0.27	72.40	4.00
IRG(Liu等，2019a)	R20	1.1	78.40	R20h	0.28	74.64	1.25
AE-KD(Du等，2020)	R56	0.85	**80.01**	R20	0.27	72.36	0.66
Self-Dist(Zhang等，2019)	R152	60	79.21	R50	25	**80.56**	2.88
ENSM(Malinin等，2019)	VGG16	140	72.50	VGG16	140	75.00	2.50

注：加粗字体表示各列最优结果，“–”表示未报告教师模型性能或不存在基线学生模型。

经过提取的此类参考信号容易进行后处理或参与额外任务，以指导学生模型训练。即使如Yim等人(2017)使用二阶项FSP矩阵匹配的方式，其特征图$\boldsymbol{v}^i$也直接来自于教师与学生的相应层级。然而，在

目标检测、实例分割和语义分割等更复杂的视觉任务中，用于进行蒸馏的模型输出不似分类向量一般容易获取。

3　检测、分割等视觉任务中的模型蒸馏

目标检测、实例分割与语义分割是应用广泛且富有挑战性的高层语义分析任务，较图像分类任务更为复杂且困难。在图像分析及语义理解领域，对图像中物体的定位问题广泛存在。目标检测将目标定位问题中的单目标推广至多目标，实例分割任务在此基础上，要求输出覆盖具体目标形状的掩膜。语义分割任务虽并无计数、定位的要求，但对相异类别物体间的边界需要特别关注。针对上述几类需要计数、分类、定位以及划分掩膜的任务，Faster R-CNN（Ren 等，2017）、RetinaNet（Lin 等，2017）、Mask R-CNN（He 等，2017）以及 DeepLab（Chen 等，2016a）等检测、分割模型的结构也较常规的分类模型复杂许多，通常包含多个不同的分支模块。

针对检测、分割任务的模型蒸馏策略进行研究需考察不同模块的独特作用。一般地，通用的检测、分割模型通常包含一个复杂的深度卷积网络 $f_b(\cdot;\boldsymbol{w}_b)$，其作为主干进行初始特征 $\boldsymbol{v}_b=f_b(\boldsymbol{I};\boldsymbol{w}_b)$ 的提取，并由后续任务分支进行处理，以获取形式各异的输出 $\boldsymbol{y}^i=f_h^i(\boldsymbol{v}_b;\boldsymbol{w}_h^i)$（上标 i 表示第 i 个任务分支）。因此，引入了针对“主干+任务分支”结构的模型蒸馏策略的探索。当前，面向检测、分割等任务的模型蒸馏工作数量相对较少，尚处于迅速发展阶段。仅针对实例分割模型的蒸馏方案更是极为罕见，现存的部分方案也是参照目标检测的模型蒸馏方式。本节介绍典型的针对目标检测、语义分割任务的模型蒸馏工作，并对其性能进行对比与分析。

3.1　目标检测中的模型蒸馏

3.1.1　针对检测器设计蒸馏策略

基于常规的“主干网络+任务分支”模型结构，针对目标检测器的蒸馏策略通常对不同特征与任务进行特殊设计。Chen 等人（2017a）针对两阶段检测器 Faster R-CNN（Ren 等，2017）的不同模块提出了蒸馏的方法，结合 Hinton 等人（2015）的基础模型蒸馏方案及 Romero 等人（2015）的线索学习策略，首先对教师与学生检测器主干网络 $f_b(\boldsymbol{I};\boldsymbol{w}_b)$ 输出的特征图 $\boldsymbol{v}_b$ 应用 L2 损失函数，使学生主干的输出特征逼近教师特征，而对任务分支中分类分支 $f_{\text{cls}}(\boldsymbol{v}_b;\boldsymbol{w}_{\text{cls}})$ 输出的分类概率向量 $\boldsymbol{p}_{\text{S}}$，则利用加权交叉熵损失函数进行蒸馏，具体为

$$L_{\text{soft}}=-\sum \boldsymbol{w}_{\text{cls}}\boldsymbol{p}_{\text{T}}\log\boldsymbol{p}_{\text{S}} \tag{29}$$

式中，$\boldsymbol{w}_{\text{cls}}$ 为类别权重，$\boldsymbol{p}_{\text{T}}$ 与 $\boldsymbol{p}_{\text{S}}$ 分别为教师及学生分类分支的输出分类向量。对于回归分支的输出 $\boldsymbol{r}_{\text{S}}=f_{\text{reg}}(\boldsymbol{v}_b;\boldsymbol{w}_{\text{reg}})$，Chen 等人（2017a）并未进行真正意义上的模型蒸馏，而是构造额外的二次优化项，将教师网络回归输出作为该优化项的松弛下界约束，即令损失函数为

$$L_{\text{reg}}=\begin{cases}\left\|\boldsymbol{r}_{\text{S}}-\boldsymbol{r}\right\|_2^2 & \left\|\boldsymbol{r}_{\text{S}}-\boldsymbol{r}\right\|_2^2+\xi>\left\|\boldsymbol{r}_{\text{T}}-\boldsymbol{r}\right\|_2^2\\ 0 & \text{其他}\end{cases} \tag{30}$$

式中，ξ 为松弛系数，$\boldsymbol{r}$ 为 ground truth 对应的回归偏移量。以此控制是否对常规回归损失项额外添加二次惩罚项。该方法在以 AlexNet（Krizhevsky 等，2017）和 VGG（Visual Geometry Group）系列（Simonyan 和 Zisserman，2015）等多种卷积网络为主干的检测器上进行实验，在 PASCOL VOC（Everingham 等，2010）和 MS COCO（Lin 等，2014）数据集上均取得了性能提升，但存在显著不足：1）在利用 L2 损失函数蒸馏主干特征图时，仅参照 Romero 等人（2015）的线索学习方案选取特征图 $\boldsymbol{v}_b$ 的完整区域，忽略了检测任务重视前景目标部分的特性；2）提出使用软标签对分类分支进行蒸馏，未提及在教师、学生 Faster R-CNN（Ren 等，2017）的 RPN（region proposal network）部分不同的情况下，若第 1 阶段的 N_{RoI} 个候选区域 $\{\boldsymbol{P}\}_{i=1}^{N_{\text{RoI}}}$ 不同，如何在两个模型间匹配区域；3）单纯将教师模型的回归结果作为训练约束条件，事实上并未使回归分支 $f_{\text{reg}}(\boldsymbol{v}_b;\boldsymbol{w}_{\text{reg}})$ 进行知识迁移。

如上文所述，检测器对前景部分局部特征关注度较高。Li 等人（2017a）尝试直接将 Hinton 等人（2015）提出的基础知识蒸馏方法应用于目标检测，但结果并不理想，因而考察典型的两阶段检测网络中，第 1 阶段初始候选框 $\{\boldsymbol{P}\}_{i=1}^{N_{\text{RoI}}}$ 在其中的影响，其以 Faster R-CNN（Ren 等，2017）及 R-FCN（region-based fully convolutional network）（Dai 等，2016b）检测器为例，利用预训练的 RPN，共用于教师与学生模型，以用于第 1 阶段感兴趣区域 $\{\boldsymbol{P}\}_{i=1}^{N_{\text{RoI}}}$ 提取；而后，使学生模型在各个区域 $\boldsymbol{P}_i$ 内均对教师模型相应区域的特

征$\boldsymbol{v}_{\mathrm{T}}[\boldsymbol{P}_i]$进行逼近，$\boldsymbol{v}[\cdot]$表示从特征中裁剪的部分（下同）。同时，参照线索学习（Romero等，2015）方案，添加了用于对齐通道数的回归层$r(\cdot)$，即令损失函数为

$$L_m = \frac{1}{2N_{\mathrm{RoI}}}\sum_i \frac{1}{m_i}\left\| \boldsymbol{v}_{\mathrm{T}}[\boldsymbol{P}_i] - r\left(\boldsymbol{v}_{\mathrm{S}}[\boldsymbol{P}_i]\right)\right\|_2^2 \tag{31}$$

式中，m_i代表各$\boldsymbol{v}_{\mathrm{S}}[\boldsymbol{P}_i]$维度大小，以作为归一化参数。对区域$\boldsymbol{P}_i$，由于分类分支$f_{\mathrm{cls}}(\boldsymbol{v}_b;\ \boldsymbol{w}_{\mathrm{cls}})$与回归分支$f_{\mathrm{reg}}(\boldsymbol{v}_b;\ \boldsymbol{w}_{\mathrm{reg}})$将分别输出分类与回归结果，因此作者亦尝试了使用L2损失进行匹配的双阶段蒸馏方案。

上述Mimic（Li等，2017a）方案利用候选提取区域$\{\boldsymbol{P}\}_{i=1}^{N_{\mathrm{RoI}}}$，保证学生与教师模型的后续任务分支$f_{\mathrm{cls}}(\boldsymbol{v}_b;\ \boldsymbol{w}_{\mathrm{cls}})$与$f_{\mathrm{reg}}(\boldsymbol{v}_b;\ \boldsymbol{w}_{\mathrm{reg}})$所处理的区域是一致的，从而准确地进行了区域匹配。然而，对常见自然图像而言，RPN提取的候选区域$\{\boldsymbol{P}\}_{i=1}^{N_{\mathrm{RoI}}}$中往往存在数量占绝对优势的大量负例，即分配至背景的区域$\{\boldsymbol{P}\}_{i=1}^{N_{\mathrm{neg}}}$，$N_{\mathrm{neg}}$为此类区域的数量。同时，此蒸馏方案中，各$\boldsymbol{P}_i$对损失函数的贡献权重均相同，导致大量负例$\{\boldsymbol{P}\}_{i=1}^{N_{\mathrm{neg}}}$携带的信息可能在蒸馏中淹没数量较少的目标前景$\{\boldsymbol{P}\}_{i=1}^{N_{\mathrm{pos}}}$中的信息。尤其对面积较小的负例候选框，其在损失函数计算中由于归一化项$1/m_i$的作用，占有与面积较大的其余区域相似的比率，这便削弱了目标前景部分的信息。

Wang等人（2019）针对两阶段的目标检测器，考察了对蒸馏主干输出的特征图$\boldsymbol{v}_b$的蒸馏方案。与Mimic（Li等，2017a）类似，利用检测器在第1阶段输出的提取区域$\{\boldsymbol{P}\}_{i=1}^{N_{\mathrm{RoI}}}$，并组成区域的并集以筛选需要进行蒸馏的特征范围，具体为

$$\boldsymbol{M} = \bigcup_{IoU>\epsilon} \boldsymbol{P}_i \tag{32}$$

式中，设定了候选区域与ground truth目标框对应区域的IoU阈值ϵ，以筛选高于该阈值的前景$\boldsymbol{P}_i$。同时，由于Faster R-CNN（Ren等，2017）等模型分配$\boldsymbol{P}_i$所对应的前景或背景时，需要参照前景与目标框的IoU，因此这种方式使前景候选框$\{\boldsymbol{P}\}_{i=1}^{N_p}$占据了$\boldsymbol{M}$中的绝大部分，而$\boldsymbol{M}$作为掩膜，便可对主干输出的特征图$\boldsymbol{v}_b$依据不同图像中不同目标的位置、大小进行选择性的蒸馏。这种方案有效选取了主干特征图$\boldsymbol{v}_b$中的前景区域，同时并未利用所有候选框对分类与回归分支的知识进行传递。与之不同，Dai等人（2021）提出了通用实例（general instance，GI）的概念，对教师与学生模型的所有前景候选区域进行筛选。计算两个模型对同一实例框输出各预测类别的置信度差异，从中选取最大的差值作为该实例框的GI置信度p_{GI}，从而筛选出教师、学生表现差异最明显的区域，用于蒸馏。蒸馏时，除对池化后的特征$\boldsymbol{v}[\boldsymbol{P}_{\mathrm{GI}}]$以及输出的分类、回归结果进行蒸馏，采用类似Park等人（2019）生成三元组的方案，对不同池化特征间的相关关系利用类似式（18）的逼近方式进行相关关系蒸馏。上述两类区域选择方法均围绕候选框进行，可能忽略若干上下文信息，生成的区域筛选掩膜为二值化掩膜，区域内不同位置的权重完全一致。

Zhang和Ma（2021）利用注意力机制对教师及学生的特征分别进行基于空间和通道的池化，以获得二者的通道注意力图与空间注意力图。蒸馏时，选取L2损失函数使教师与学生的两类注意力图进行逼近。同时，将两个模型的注意力图合并，利用softmax运算获得归一化通道掩膜$\boldsymbol{M}^c$与空间掩膜$\boldsymbol{M}^s$，以此构造注意力引导的2范数蒸馏损失，即

$$L_{\mathrm{AM}} = \left\| \boldsymbol{M}^c \cdot \boldsymbol{M}^s \cdot \left(\boldsymbol{v}_{\mathrm{S}} - \boldsymbol{v}_{\mathrm{T}}\right)\right\|_2 \tag{33}$$

式中，运算符“$\cdot$”代表逐点相乘的掩膜操作。同时，通过非局部（non-local）模块计算主干特征$\boldsymbol{v}_{\mathrm{S}}$与$\boldsymbol{v}_{\mathrm{T}}$的非局部关系矩阵$\boldsymbol{A}_{\mathrm{S}}$与$\boldsymbol{A}_{\mathrm{T}}$，使二者逼近，以利用非局部蒸馏传递不同位置的特征间关系。这种方法基于特征注意力与非局部运算，自适应地将特征间、区域间的重要性与相互关系充分利用。然而，依此选取的蒸馏区域相对缺少目标区域等语义解释。

3.1.2 特殊检测任务中应用模型蒸馏

除了针对检测器设计蒸馏策略，检测器蒸馏的研究也涉及蒸馏方法在基于目标检测的其他问题中的应用。Mehta和Ozturk（2018）考察YOLO（Redmon等，2016）等单阶段检测器的实时检测任务时，采用更轻量且推断速度更快的学生检测器Tiny YOLO对教师模型进行替换的压缩方案，引入模型蒸馏方法以提升轻量检测器性能。由于YOLO检测器未引入锚框概念，需预先设定预测的目标框数量，因此蒸馏方式相对简单，对教师、学生检测器各设定区域$\boldsymbol{P}_i$（例如标准YOLO检测器中划分的每个矩形方格）中的对应目标框$\boldsymbol{B}_{\mathrm{gt}}$，将教师的输出目标$\boldsymbol{B}_{\mathrm{T}}$作为学生训

练中额外的 ground truth 进行监督即可。考虑到添加教师网络检测结果会造成边界框的大量重复，专门设计了特征图非极大值抑制（feature map-non maximum suppression，FM-NMS），在蒸馏前筛选高置信度的教师模型输出作为软标签，使目标状态（是否前景）$f_{\text{obj}}^{\text{Comb}}$、分类结果 $f_{\text{cls}}^{\text{Comb}}$ 和边界框位置 $f_{\text{reg}}^{\text{Comb}}$ 三部分共同作为新的 ground truth 进行蒸馏。以此，其学生模型可在超过 200 幅/s 的推断速度下大幅缩小与教师模型的性能差距。

弱监督目标检测（weakly supervised object detection，WSOD）通常仅存在图像级别的多类别分类标注 $\{\boldsymbol{y}_i\}_{i=1}^{N_{\text{cls}}}$（$N_{\text{cls}}$ 为类别数量），因此利用类似模型自蒸馏的方案，基于模型自身输出伪标签 $\boldsymbol{B}_{\text{T}}$ 进行二次训练是常用方法。Zeng 等人（2019）采样初始阶段网络对各个提取区域的预测结果 $\boldsymbol{p}_{\text{S}}$，首先对其进行 NMS，然后考察特征图基于类别与区域间 softmax 归一化的置信度筛选，最后选取临近区域对应的检测输出 $\hat{\boldsymbol{p}}_{\text{S}}$ 作为下一轮迭代的软标签。Huang 等人（2020）则面向不同的感兴趣区域（region of interest，RoI）池化后的特征 $\{\boldsymbol{v}_b[\boldsymbol{P}_i]\}_{i=1}^{N}$，将 K 个不同变换的输出 $\{T_k(\boldsymbol{v}_b[\boldsymbol{P}_i])\}_{k=1}^{K}$ 与 N_l 个不同后续网络层的输出 $\{\boldsymbol{v}_b[\boldsymbol{P}_i]\}_{i=1}^{N_l}$ 分别提取，并利用逐通道的平均值组合为注意力图，具体为

$$\boldsymbol{M}_{\text{att}} = \text{sigmoid}\left(\frac{1}{C}\sum_{l_c=1}^{C}\boldsymbol{v}_b[\boldsymbol{P}_i]^{l_c}\right) \tag{34}$$

式中，C 为当前特征通道数，$\boldsymbol{v}_b[\boldsymbol{P}_i]^{l_c}$ 表示第 l_c 通道的特征。基于 $\boldsymbol{M}_{\text{att}}$ 逐点取极大值，生成逐实例的注意力图 $\boldsymbol{M}_{\text{att}}^{\text{IW}}$ 与逐层的注意力图 $\boldsymbol{M}_{\text{att}}^{\text{LW}}$，与 $\{T_k(\boldsymbol{v}_b[\boldsymbol{P}_i])\}_{k=1}^{K}$ 共同构造后续迭代过程中的蒸馏监督信号。

3.1.3　检测器蒸馏方法的性能比较

前文提到，相比于成熟的分类器蒸馏方法，针对目标检测任务的检测器蒸馏方案较为稀少。本文在调研工作中发现，相同的模型由于训练细节的不同，其报告的基线性能可能也不相同，因而相同的学生模型经过蒸馏，性能可能存在差异。因此，本文尽可能依照模型与数据集的一致性选取各方案中的部分实验结果进行展示及分析，以保证性能对比足够公平。

本文选取典型目标检测数据集 MS COCO 2017（Lin 等，2014）的验证集和 PASCAL VOC 2007（Everingham 等，2010）的测试集作为统一测试集，各方案实验性能对比结果如表 5 和表 6 所示。其中，教师模型与蒸馏后学生模型的性能基于数据集存在不同指标，MS COCO（Lin 等，2014）使用 AP50∶95，PASCAL VOC 使用 AP50，两个指标的含义在 1.1 节已有介绍，数值均保留 1 位小数；性能提升指相对基线模型（如果有）的性能增加值。学生与教师模型栏与表 2—表 4 类似，含义如下：1）写明模型主干的，如 R50∶23，表示使用 Faster R-CNN（Ren 等，2017）检测器，主干网络简写含义与表 2—表 4 基本一致，后缀字母 h 表示参数量削减一半获得的轻量主干网络。2）写明检测器类型的，如 Cascade，YOLO 等，表示 Cascade R-CNN（Cai 和 Vasconcelos，2018）与 YOLO（Redmon 等，2016）系列检测器及其变体，F-YOLO 为基于 TinyYOLO 蒸馏获得的轻量级检测器。3）参数量单位仍为 M，由于多数检测模型蒸馏算法均在相同检测模型类型的条件下替换主干网络，因此此类方案的参数量代表主干网络参数量，不包含任务分支。OD200F（object detection at 200 frames per second）（Mehta 和 Ozturk，2018）方案使用不同结构的教师及学生检测器，参数量表示检测器整体参数量。各参数量数值保留两位有效数字。

此外，部分方法未报告教师检测器性能，表中以“-”表示；部分任务属于弱监督目标检测任务（仅含图像层级多类别标注，无目标框），设计了特殊结构检测器，用以进行自蒸馏，其性能相对全监督检测任务较低，且不存在教师模型性能及基线学生模型性能，也以“-”表示。各类网络变体的含义及结构详见参考文献。

观察表 5 和表 6 能够发现，对于蒸馏后的性能，自身结构更复杂的学生检测器更易获得较高的 AP。例如，在两个数据集上，最终性能表现最好的模型主干均为 R50（ResNet-50（He 等，2016）），而基线性能较低的学生模型往往更容易获得较高的提升。本身性能较高的学生检测器，如 GISM（general instance selection module）（Dai 等，2021）在 PASCAL VOC 2007 数据集上 AP50 达到 82.6 的 R50，仅获得 0.4 的微小性能提升。这是由于对规模较小的数据集，带有复杂主干的检测器已经逼近性能上限，同时教师、学生之间的性能差距相对较小。上述两点与对表 2—表 4 的分析结论基本一致。总体而言，尽管各

类模型蒸馏方法均可成功提升目标检测模型性能，但选取的教师、学生检测器种类、结构各异，实验配置也存在不同，方法缺乏通用性。

对蒸馏策略本身而言，用于目标检测器的模型蒸馏大多关注对特征中有效局部信息尤其是前景信息的增强。事实上，选取特征中对学生模型较为重要的区域进行蒸馏尝试，例如使用注意力机制、梯度等对不同区域的特征进行加权，正是目前对检测器特征进行更精细利用的策略。未来研究中，对不同区域特征的信息提取与迁移仍是检测器蒸馏的重点。

表5　在MS COCO 17数据集上，针对目标检测器的部分模型蒸馏方法的实验性能

Table 5　Experiment results of several model distillation methods for object detectors on MS COCO 17 dataset

蒸馏方法	教师模型			学生模型			ΔAP/%
	类型	参数量/M	AP50:95/%	类型	参数量/M	AP50:95/%	
FGFI(Wang等，2019)	R50	23	36.9	R50h	5.7	34.8	**3.6**
FBKD(Zhang和Ma，2021)	Cascade X101	40	–	R50	23	**41.5**	3.1
GISM(Dai等，2021)	R101	42	**38.3**	R50	23	40.2	1.9
EfficientOD(Chen等，2017a)	VGG16	15	24.2	VGGM	6.5	17.3	1.2
WSOD2(Zeng等，2019)	VGG16	15	–	VGG16	15	10.8	–
CASD(Huang等，2020)	VGG16	15	–	VGG16	15	12.8	–

注：加粗字体表示各列最优结果，“–”表示未报告教师模型性能或不存在教师模型或基线学生模型性能。

表6　在PASCAL VOC 07数据集上，针对目标检测器的部分模型蒸馏方法的实验性能

Table 6　Experiment results of several model distillation methods for object detectors on PASCAL VOC 07 dataset

蒸馏方法	教师模型			学生模型			ΔAP/%
	类型	参数量/M	AP50/%	类型	参数量/M	AP50/%	
Mimic(Li等，2017a)	R50	23	–	VGG16h	3.7	48.7	5.2
OD200F(Mehta和Qzturk，2018)	YOLOv2	51	73.4	F-YOLO	7.9	66.9	7.5
GISM(Dai等，2021)	R101	42	**82.8**	R50	23	**82.6**	0.4
FGFI(Wang等，2019)	VGG16	15	70.4	VGG11	10	67.6	**8.0**
EfficientOD(Chen等，2017a)	VGG16	15	70.4	VGGM	6.5	63.7	3.9
WSOD2(Zeng等，2019)	VGG16	15	–	VGG16	15	56.0	–
CASD(Huang等，2020)	VGG16	15	–			56.8	–

注：加粗字体表示各列最优结果，“–”表示未报告教师模型性能或不存在教师模型或基线学生模型性能。

3.2　语义分割中的模型蒸馏

由于目标检测与实例分割存在紧密联系，而专门针对实例分割的模型蒸馏工作尚为鲜见，因此本节主要介绍模型蒸馏方法在语义分割任务中的应用。相比于目标检测与实例分割重视目标定位、大小的特点，语义分割并不集中关注局部区域，而是考察特征全局像素点以及不同区域像素的关系，这也是对语义分割器进行模型蒸馏的关注点。

针对语义分割的模型蒸馏研究较检测器蒸馏的有关研究更为稀缺，本节依次介绍典型的数种方案，对其性能进行简单的比较与分析。

3.2.1　分割器蒸馏基础方法：模块化蒸馏

由于语义分割模型通常包含显著的模块化结构：主干（编码器）、重建层（译码器）和softmax输出层等，对分割器的基础蒸馏方法便建立于其不同模块的输出特征上。Liu等人（2019b）的研究为典型模块化蒸馏的方案。首先，对深度模型输出分割掩膜$\boldsymbol{M}$以及主干输出特征图$\boldsymbol{v}$进行处理。对于前者，仿照对分类器蒸馏的一般方式，对各像素点对应的分类向量使用KL散度进行逼近，即

$$L_{pi}=\frac{1}{W'\times H'}\sum_{i\in\boldsymbol{R}}\mathrm{KL}(\boldsymbol{M}_{\mathrm{S}}^{i}\|\boldsymbol{M}_{\mathrm{T}}^{i}) \tag{35}$$

式中，W'与H'分别为掩膜的宽与高，$\boldsymbol{R}$为掩膜区域，$\boldsymbol{M}_{\mathrm{S}}^{i}$与$\boldsymbol{M}_{\mathrm{T}}^{i}$分别表示教师与学生模型输出掩膜于$\boldsymbol{R}$中采样的像素；对输出特征图$\boldsymbol{v}$，计算逐通道归一化相似度矩阵，获得由余弦相似度组成的矩阵$\boldsymbol{A}$，再利用L2损失函数进行蒸馏。其次，与分类任务中Xu等人（2018）的研究类似，引入对抗判决的思想，在教师与学生分割器$S(\cdot)$的末端接入对抗判决器$D(\cdot)$，使教师及学生模型输出的分割结果与原图像$\boldsymbol{I}$共同输入判决器，分别输出教师与学生的嵌入（embedding），再基于Wasserstein距离计算损失，并作为整体性损失函数在蒸馏过程中交替进行蒸馏学生模型、训练判决器的流程。具体为

$$L_{ho}(S,D)=E_{q_{\mathrm{S}}}\left[D\left(\boldsymbol{M}_{\mathrm{S}}|\boldsymbol{I}\right)\right]-E_{q_{\mathrm{T}}}\left[D\left(\boldsymbol{M}_{\mathrm{T}}|\boldsymbol{I}\right)\right] \tag{36}$$

式中，$E_q[\cdot]$为基于q的分布求期望，$D\left(\boldsymbol{M}|\boldsymbol{I}\right)$为判决器在输入图像$\boldsymbol{I}$的条件下输出的概率。利用上述“蒸馏+对抗”的方案，学生模型在Cityscapes（Cordts等，2016）等数据集上获得了显著性能提升。

类似利用相似度、对抗方法辅助语义分割器蒸馏的策略也为研究人员沿用。在Wang等人（2020c）算法中，针对模型输出的分割掩膜$\boldsymbol{M}$进行蒸馏，而于模型全局仍采用交叉熵损失、对抗判决器以及相似度的L2损失进行训练。关键改进在于构造了一张新的掩膜，称为类内特征变化图（intra-class feature variation，IFV），具体为

$$\boldsymbol{M}^{j}=sim\left(\boldsymbol{M}[j],\frac{1}{|\boldsymbol{S}|}\sum_{i\in\boldsymbol{S}}\boldsymbol{M}[i]\right) \tag{37}$$

式中，i与j均为来自同一类别像素集合$\boldsymbol{S}$的不同像素坐标索引，$sim(\cdot,\cdot)$为余弦相似度。基于IFV，将余弦相似度计算限制在各相同类别像素点之间，而后使教师与学生模型的IFV进行逼近以进行蒸馏。这种方案使知识的迁移能够清晰地在各类别的内部进行。

不同于直接选取各网络模块输出的方案，He等人（2019）沿用了分类与检测任务的模型蒸馏中对特征图进行后处理再蒸馏的思路，训练一个额外的自编码器$D_c\left(E_c\left(\boldsymbol{v}_{\mathrm{T}}\right)\right)$对教师网络主干输出的特征进行重建，从而在编码器的输出中获得更为“紧致的知识”$E_c\left(\boldsymbol{v}_{\mathrm{T}}\right)$用以学生模型逼近。同时，学生模型选取包含批标准化的卷积层$Conv_{\mathrm{BN}}\left(\boldsymbol{v}_{\mathrm{S}}\right)$的输出以进行对齐。最后，使归一化的此两类特征基于某特定范数计算损失函数。另一方面，He等人（2019）采用考察像素间相似度的方案，利用后处理层$\boldsymbol{A}(\cdot)$对上述特征计算仿射矩阵$\boldsymbol{A}_{\mathrm{S}}\left(Conv_{\mathrm{BN}}\left(\boldsymbol{v}_{\mathrm{S}}\right)\right)$与$\boldsymbol{A}_{\mathrm{T}}\left(E_c\left(\boldsymbol{v}_{\mathrm{T}}\right)\right)$后，再基于2范数距离使二者进行逼近。上述两种损失函数共同对学生模型进行蒸馏，参与计算的元素并非直接取自教师及学生模型的输出特征，而均来自二者的不同后处理输出。

3.2.2　基于类间、像素间关系的分割器蒸馏

上一小节介绍的分割器蒸馏方案均基于基础的模块化蒸馏，对不同模块结构的输出进行逼近。而类似2.1.5节中基于样本间关系的分类器蒸馏方案，研究人员同样利用分割器的输出构造了隐式的相关关系进行蒸馏。与分类器不同，语义分割关注图像全局，考察不同位置的像素分类，因此分割器的蒸馏通常考察类间、像素间关系。Xie等人（2018）与Shu等人（2021）均对特征的逼近方式进行改进，在蒸馏前对主干输出特征图$\boldsymbol{v}$以及分割掩膜$\boldsymbol{M}$进行分析，以考察像素或类别间的关系。其中，Xie等人（2018）利用$\boldsymbol{M}$构造两个用于逼近的特征，其中一个在各像素进行softmax归一化，获得一幅概率特征图，另一个则根据各像素的8-邻接像素生成“一致性图”，即

$$\boldsymbol{M}^{c}(v)=\sum_{u\in\boldsymbol{R}_B(v)}\left\|\boldsymbol{M}(u)-\boldsymbol{M}(v)\right\|_2^2 \tag{38}$$

式中，v为掩膜中的某个像素，而$\boldsymbol{R}_B(v)$为临近该像素的8–邻接像素区域。依此构造的$\boldsymbol{M}^c$类似于对$\boldsymbol{M}$使用拉普拉斯算子进行处理的结果，有效提取了分割掩膜边界的信息。该方法使学生与教师的此两类特征进行逼近，以同时顾及分类信息与分割边界信息的迁移，达到蒸馏目的。与之不同，后者（Shu等，2021）同时选取$\boldsymbol{v}$与$\boldsymbol{M}$进行归一化，不采用进行逐像素归一化的方案，而在各通道内对所有像素进行逐通道softmax归一化，随后进行逼近。该逐通道归一化策略考察了不同通道的像素间关系，而非使学生与教师模型独立地在像素所在位置进行逼近。

相较之下，Park和Heo（2020）对特征图不同通道的相互关系进行了更充分的利用。其基于张量乘积，首先将特征$\boldsymbol{v}\in\mathbf{R}^{C\times H\times W}$依各通道顺序进行循环移位，并与$\boldsymbol{v}$本身逐点相乘获得$C$个自相关张量，再将其合并，并展平为$HW\times C^2$的2维自相关矩阵$\boldsymbol{A}_{\mathrm{cor}}$，最后利用矩阵乘积，生成表达跨通道相关性的矩阵$\boldsymbol{A}_{\mathrm{sim}}=\boldsymbol{A}_{\mathrm{cor}}\boldsymbol{A}_{\mathrm{cor}}^{\mathrm{T}}$后，再利用归一化L2损失使学生与教师的$\boldsymbol{A}_{\mathrm{sim}}$进行逼近。同时，将分类掩膜中的概率向

量利用类似式(23)的混合方法生成一个可控制“软化程度”的软标签掩膜,而后再用于蒸馏。此方案的亮点在于对不同通道的特征相互关系的考察。然而,通过循环移位以进行跨通道的自相关计算忽略了通道之间存在的分布差异。而针对特征表示相对独立的各通道计算其自相关张量缺少可靠的解释。

3.2.3 语义分割器蒸馏方法的性能对比

针对语义分割的模型蒸馏方法较检测器蒸馏方案更加稀少,因此本文继续依照“数据集尽量统一、模型尽量相似”的原则选取所调研方法的部分实验结果进行展示与公平对比。

本文选取典型的Cityscapes数据集(Cordts等,2016)中包含完整语义分割标注的测试集,以及PASCAL VOC 2012数据集(Krizhevsky和Hinton,2009)中带有语义分割标注的验证集作为性能对比的通用测试集。表7和表8展示了不同分割器蒸馏方案的性能。其中,教师与学生模型性能统一以mIoU衡量,性能提升指该指标的提升值,上述指标均保留1位小数。表中教师与学生模型栏的简写代表分割模型的主干网络,XC代表Xception(Chollet,2017)网络,MobileV2与MobileV1分别指MobileNetV2(Sandler等,2019)及MobileNetV1(Howard等,2017),其余简写含义与表2—表4一致。为尽可能进行公平比较,本文选择基于相似主干的蒸馏结果进行展示。与表5和表6类似,参数量保留两位有效数字,均表示主干本身参数量,与语义分割模型的反卷积、上采样等特征处理分支无关。关于具体分割模型,KA(knowledge adaptation)(He等,2019)仅单纯利用主干特征生成分割掩膜,TSL(teacher-student learning)(Xie等,2018)选取DeepLabV2(Chen等,2016a)分割模型,CSCACE(channel and spatial correlations and adaptive cross entropy)(Park和Heo,2020)选取DeepLabV3+(Chen等,2018)分割模型,其余方案均选取PSPNet(pyramid scene parsing network)(Zhao等,2017)作为应用各类主干的语义分割器。

从表7和表8可观察到,3项选取PSPNet分割器且使用ResNet-101主干对ResNet-18主干进行蒸馏的研究中,CWD(channel-wise distillation)(Shu等,2021)在Cityscapes的测试集上取得了最高的绝对性能与最大的性能提升。一方面由于其通道内归一化使得进行逼近的张量缩放至相似的尺度(0~1),另一方面由于这种归一化同时考察了特征与分割掩膜二者不同位置像素之间的相关关系。由于实验配置的差异,这3种方案基线性能并不完全相同,而其对性能提升的公平比较并无太大影响。其中,CSCACE(Park和Heo,2020)获得了最大的性能提升值,而其绝对性能的显著偏低是其较低的基线性能导致的,实验证明并不影响对学生模型的性能提升能力。在PASCAL VOC 12的验证集上,IFVD(intra-class feature variation distillation)(Wang等,2020c)取得了最高的mIoU;同时CWD(Shu等,2021)仍然在性能提升上取得了良好效果。说明在对语义分割器进行蒸馏时,更细致地考察各像素类别、不同位置像素间的关系可显著改善模型蒸馏性能。

总体上,由于语义分割更多考察特征图上逐像素的向量以及不同像素的分类输出,因此针对语义分割器的模型蒸馏也大多基于分割掩膜$\boldsymbol{M}$、主干输出特征$\boldsymbol{v}$和相似度矩阵或张量三方面输出进行逼近。未来研究中,一方面可能出现构造更合适相似度的方案,如参考2.1.5中各类利用不同相关性度量的方案;另一方面可能与TSL(Xie等,2018)类似,在蒸馏中对部分能够显著影响分割性能的区域,例

表7 在Cityscapes测试集上,针对语义分割模型的部分模型蒸馏方法的实验性能

Table 7 Experiment results of several model distillation methods for semantic segmentation models on Cityscapes test set

蒸馏方法	教师模型			学生模型			ΔmIoU/%
	类型	参数量/M	mIoU/%	类型	参数量/M	mIoU/%	
KA(He等,2019)	XC41	25	–	MobileV2	1.3	72.7	2.5
CSCACE(Park和Heo,2020)	XC65	38	72.6	R18	9.6	69.7	**6.6**
IFVD(Wang等,2020c)	R101	42	**78.6**	R18	9.6	72.7	5.1
CWD(Shu等,2021)	R101	42	78.5	R18	9.6	**74.6**	7
SKD(Liu等,2019b)	R101	42	78.4	R18	9.6	71.4	3.8

注:加粗字体表示各列最优结果。

表 8　在 PASCAL VOC 12 验证集上，针对语义分割模型的部分模型蒸馏方法的实验性能

Table 8　Experiment results of several model distillation methods for semantic segmentation models on PASCAL VOC 12 validation set

蒸馏方法	教师模型			学生模型			ΔmIoU/%
	类型	参数量/M	mIoU/%	类型	参数量/M	mIoU/%	
KA(He 等,2019)	R50	23	76.2	MobileV2	1.3	72.5	1.9
TSL(Xie 等,2018)	R101	42	75.2	MobileV1	3.2	69.6	2.3
IFVD(Wang 等,2020c)	R101	42	77.8	R18	9.6	**74.1**	3.2
CWD(Shu 等,2021)	R101	42	**78.5**	–	–	69.3	**3.9**

注：加粗字体表示各列最优结果，“–”表示原文献未提供结果。

如边界附近的特征或分割掩膜，进行特别加强。

4　结　语

模型蒸馏算法提出以来，在典型的高层语义分析任务上取得了诸多进展，本文考察了针对图像分类、目标检测/实例分割以及语义分割的各类模型蒸馏算法，展示了部分蒸馏策略带来的性能提升对比。整体上，模型蒸馏的方式与应用范围更为多样，而有关模型蒸馏的研究仍然存在部分缺陷与不足。本节中，本文将简要分析当下模型蒸馏存在的问题，并针对这些问题展望其未来可能的研究方向。

1)针对分类器的模型蒸馏方法，新的知识迁移方式在性能与通用性上欠佳。因此，未来对新方案进行开发，仍将侧重于使用对抗网络、互信息建模等新的监督方法，同时应当寻找对不同结构分类网络更通用的蒸馏方案。一方面，较为成熟、性能较良好的方案往往如 2.1.4 节和 2.1.5 节描述的蒸馏方式，试图构造新的蒸馏的目标，或是对特征做合适的后处理以充分利用。例如，考察特征的相似度矩阵，对不同层、不同样本间的特征进行匹配等。然而，这些方案本质上均为对网络输出特征进行某种变换后，再进行直接逼近。另一方面，如 2.1.2，2.1.3 及 2.1.6 节的方法则与基于输出逼近的蒸馏策略不同，其中许多方案在建立新的蒸馏框架、提供新颖知识迁移方式的同时，增大了算法复杂度，而其性能提升比其余方案相较并不占优。部分新颖方案如“信息流”(Passalis 等，2020)等，实现条件限制较多，缺少在一般网络、数据集上的实践。因此，研究人员一方面需要寻找新的知识迁移方式，以替代传统知识蒸馏中“输入输出+计算损失”的框架；另一方面，应用模型蒸馏时，需要建立不同结构教师与学生模型间更通用的监督关系描述。例如，蒸馏时，考察教师模型不同层分别与学生模型对应层的相关性度量等。

2)当前的模型蒸馏任务缺少规范而通用的比较框架。因此，未来需要基于常用的分类模型与数据集，在固定有关实验配置的条件下，建立较为统一的模型蒸馏评估指标。第 2、3 节中，尽管本文在表 2—表 8 内展示了基于相同的数据集、模型条件下不同蒸馏方案的性能，但是其中的许多方案由于通用性不足，仍然对数据集选取、数据增强方式、训练迭代长度等设置并不相同。因此导致相同的基线模型(学生)性能存在差异，为绝对性能的比较引入困难。例如，自蒸馏任务通常需要与模型自身的常规训练结果进行对比，但如 ResDist(Li 等，2020a)等方法的应用模型较为特殊，无法与其余的自蒸馏方法选择相同模型进行比较。因此，未来针对分类器的模型蒸馏研究，亟待建立较为普适的模型蒸馏性能评估标准，使其能够基于通用数据集(如 ImageNet、CIFAR 系列)选择特定通用模型(如 ResNet 系列、VGG 系列)对性能提升的作用进行验证。另外，应当将常见的轻量网络如 MobileNet(Howard 等，2017)等作为学生模型，评估蒸馏对模型压缩的促进作用。在蒸馏时，也需寻求相同的基线模型、训练迭代长度以及图像预处理方案。同时，需要将蒸馏算法的复杂度以及模型的压缩加速效果纳入蒸馏性能的评估指标，对不同方法进行比较与分析。

3)针对检测、分割任务的模型蒸馏方法仍然不够完善。因此，应当考察不同任务对模型的要求，对不同结构的检测器与分割器分别设计特殊蒸馏方法。与目前结构较为统一的分类器不同，近年来针

对检测、分割任务的新颖模型之间不存在统一的结构。而目前许多相关的模型蒸馏方案将面向分类器的蒸馏方法直接迁移至下游任务，忽视了各类下游任务的特殊性。在目标检测中，模型蒸馏对重要的前景信息和目标上下文的关注度需要提升，同时应尽量抑制无关的背景噪声；在实例分割中，仅有的模型蒸馏方案仍然基于基本的Mask R-CNN（He等，2017）系列模型进行蒸馏，并对检测任务存在较大依赖；在语义分割任务中，由于需要考察不同像素点之间的关系，现存方案大多为“多点匹配特征+相似度矩阵”模式，未来可能在构造新的相似度关系、利用分割类别边界以及蒸馏语义分割器中间层特征等方面进行探索。总体上，针对目标检测、实例分割和语义分割任务，当前需要引用分类器蒸馏算法作为基础，针对不同任务与模型结构探索特定蒸馏策略。

本文主要调研应用模型蒸馏的典型任务，包括分类、检测与分割，对许多更复杂的计算视觉任务如视频目标跟踪、重识别等问题均具备支撑作用。对这几类典型任务中的模型蒸馏方案进行开发，也能够使性能获得提升的深度模型对其他计算机视觉任务提供良好支持。

参考文献（References）

Ahn S S, Hu S X, Damianou A, Lawrence N D and Dai Z W. 2019. Variational information distillation for knowledge transfer//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9163-9171 [DOI: 10.1109/CVPR.2019.00938]

Anil R, Pereyra G, Passos A, Ormandi R, Dahl G E and Hinton G E. 2020. Large scale distributed neural network training through online distillation [EB/OL]. [2020-08-20]. https://arxiv.org/pdf/1804.03235.pdf

Ba L J and Caruana R. 2014. Do deep nets really need to be deep?//Proceedings of the Advances in Neural Information Processing Systems 27. Montréal, Canada: MIT Press: 2654-2662

Badrinarayanan V, Kendall A and Cipolla R. 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12): 2481-2495 [DOI: 10.1109/TPAMI.2016.2644615]

Buciluǎ C, Caruana R and Niculescu-Mizil A. 2006. Model compression//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA: ACM: 535-541 [DOI: 10.1145/1150402.1150464]

Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: delving into high quality object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6154-6162 [DOI: 10.1109/CVPR.2018.00644]

Chen G B, Choi W G, Yu X, Han T and Chandraker M. 2017a. Learning efficient object detection models with knowledge distillation//Proceedings of the Advances in Neural Information Processing Systems 30. Long Beach, USA: Curran Associates Inc.: 742-751

Chen L C, Papandreou G, Kokkinos I, Murphy K and Yuille A L. 2016a. Semantic image segmentation with deep convolutional nets and fully connected CRFs [EB/OL]. [2016-06-07]. https://arxiv.org/pdf/1412.7062.pdf

Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 801-818 [DOI: 10.1007/978-3-030-01234-2_49]

Chen T Q, Goodfellow I and Shlens J. 2016b. Net2Net: accelerating learning via knowledge transfer [EB/OL]. [2016-04-23]. https://arxiv.org/pdf/1511.05641.pdf

Chen Y T, Wang N Y and Zhang Z X. 2017b. DarkRank: accelerating deep metric learning via cross sample similarities transfer [EB/OL]. [2017-12-18]. https://arxiv.org/pdf/1707.01220.pdf

Chollet F. 2017. Xception: deep learning with depthwise separable convolutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1251-1258 [DOI: 10.1109/CVPR.2017.195]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The cityscapes dataset for semantic urban scene understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3213-3223 [DOI: 10.1109/CVPR.2016.350]

Crowley E J, Gray G and Storkey A J. 2018. Moonshine: distilling with cheap convolutions//Proceedings of the Advances in Neural Information Processing Systems 31. Montréal, Canada: Curran Associates Inc.: 2893-2903

Dai J F, He K M, Li Y, Ren S Q and Sun J. 2016a. Instance-sensitive fully convolutional networks//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 534-549 [DOI: 10.1007/978-3-319-46466-4_32]

Dai J F, Li Y, He K M and Sun J. 2016b. R-FCN: object detection via region-based fully convolutional networks [EB/OL]. [2016-06-21]. https://arxiv.org/pdf/1605.06409.pdf

Dai J F, Qi H Z, Xiong Y W, Li Y, Zhang G D, Hu H and Wei Y C. 2017. Deformable convolutional networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 764-773 [DOI: 10.1109/ICCV.2017.89]

Dai X, Jiang Z R, Wu Z, Bao Y P, Wang Z C, Liu S and Zhou E J. 2021. General instance distillation for object detection [EB/OL]. [2021-03-03]. https://arxiv.org/pdf/2103.02340.pdf

Deng J，Dong W，Socher R，Li L J，Li K and Li F F. 2009. ImageNet：a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami，USA：IEEE：248-255［DOI：10.1109/CVPR. 2009. 5206848］

Du S C，You S，Li X J，Wu J L，Wang F，Qian C and Zhang C S. 2020. Agree to disagree：adaptive ensemble knowledge distillation in gradient space//Proceedings of the Advances in Neural Information Processing Systems 33. Virtual：Curran Associates Inc.：12345-12355

Everingham M，Van Gool L，Williams C K I，Winn J and Zisserman A. 2010. The pascal visual object classes（VOC）challenge. International Journal of Computer Vision，88（2）：303-338［DOI：10. 1007/s11263-009-0275-4］

Feng X，Jiang Y N，Yang X J，Du M and Li X. 2019. Computer vision algorithms and hardware implementations：a survey. Integration，69：309-320［DOI：10.1016/j.vlsi.2019.07.005］

Girdhar R，Tran D，Torresani L and Ramanan D. 2019. DistInit：learning video representations without a single labeled video//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea（South）：IEEE：852-861［DOI：10.1109/ICCV.2019.00094］

Gupta S，Hoffman J and Malik J. 2016. Cross modal distillation for supervision transfer//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：IEEE：2827-2836［DOI：10.1109/CVPR.2016.309］

Han S，Mao H Z and Dally W J. 2016. Deep compression：compressing deep neural networks with pruning，trained quantization and Huffman coding［EB/OL］.［2016-02-15］. https://arxiv.org/pdf/1510.00149.pdf

Han S，Pool J，Tran J and Dally W J. 2015. Learning both weights and connections for efficient neural network//Proceedings of the Advances in Neural Information Processing Systems 28. Montréal，Canada：MIT Press：1135-1143

Hariharan B，Arbeláez P，Bourdev L，Maji S and Malik J. 2011. Semantic contours from inverse detectors//Proceedings of 2011 International Conference on Computer Vision. Barcelona，Spain：IEEE：991-998［DOI：10.1109/ICCV.2011.6126343］

He K M，Gkioxari G，Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice，Italy：IEEE：2961-2969［DOI：10.1109/ICCV. 2017.322］

He K M，Zhang X Y，Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：IEEE：770-778［DOI：10.1109/CVPR.2016.90］

He T，Shen C H，Tian Z，Gong D，Sun C M and Yan Y L. 2019. Knowledge adaptation for efficient semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：578-587［DOI：10.1109/CVPR.2019.00067］

Hinton G，Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network［EB/OL］.［2015-03-09］. https://arxiv.org/pdf/1503.02531.pdf

Howard A G，Zhu M L，Chen B，Kalenichenko D，Wang W J，Weyand T，Andreetto M and Adam H. 2017. MobileNets：efficient convolutional neural networks for mobile vision applications［EB/OL］.［2017-04-17］. https://arxiv.org/pdf/1704.04861.pdf

Huang T S. 1996. Computer Vision：Evolution and Promise. CERN European Organization for Nuclear Research-Reports-CERN：21-26

Huang Z Y，Zou Y，Kumar B V K V and Huang D. 2020. Comprehensive attention self-distillation for weakly-supervised object detection//Proceedings of the Advances in Neural Information Processing Systems 33. Virtual：Curran Associates Inc.：16797-16807

Jiang W，Chan K L，Li M J and Zhang H J. 2005. Mapping low-level features to high-level semantic concepts in region-based image retrieval//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego，USA：IEEE：244-249［DOI：10.1109/CVPR.2005.220］

Jin X，Peng B Y，Wu Y C，Liu Y，Liu J H，Liang D，Yan J J and Hu X L. 2019. Knowledge distillation via route constrained optimization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul，Korea（South）：IEEE：1345-1354［DOI：10.1109/ICCV.2019.00143］

Kim Y D，Park E H，Yoo S J，Choi T L，Yang L and Shin D J. 2016. Compression of deep convolutional neural networks for fast and low power mobile applications［EB/OL］.［2016-02-24］. https://arxiv.org/pdf/1511.06530.pdf

Krasin I，Duerig T，Alldrin N，Ferrari V，Abu-El-Haija S，Kuznetsova A，Rom H，Uijlings J，Popov S and Kamali S. 2018. OpenImages：a public dataset for large-scale multi-label and multi-class image classification［DB/OL］.［2018-05-01］. https://github.com/openimages

Krizhevsky A and Hinton G. 2009. Learning multiple layers of features from tiny images［EB/OL］.［2009-04-08］. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf

Krizhevsky A，Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM，60（6）：84-90［DOI：10.1145/3065386］

LeCun Y，Bottou L，Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE，86（11）：2278-2324［DOI：10.1109/5.726791］

LeCun Y，Cortes C and Burges C J C. 2013. The MNIST database of handwritten digits［DB/OL］.［2013-05-14］. http://yann.lecun.com/exdb/mnist/

Li G L，Zhang J L，Wang Y H，Liu C J，Tan M，Lin Y F，Zhang W，Feng J S and Zhang T. 2020a. Residual distillation：towards por-

table deep neural networks without shortcuts//Proceedings of the Advances in Neural Information Processing Systems 33. Virtual: Curran Associates Inc.: 8935-8946

Li Q Q, Jin S Y and Yan J J. 2017a. Mimicking very efficient network for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6356-6364 [DOI: 10.1109/CVPR.2017.776]

Li T H, Li J G, Liu Z and Zhang C S. 2020b. Few sample knowledge distillation for efficient network compression//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE: 14639-14647 [DOI: 10.1109/CVPR42600.2020.01465]

Li W, Wang L M, Li W, Agustsson E and Van Gool L. 2017b. WebVision database: visual learning and understanding from web data [EB/OL]. [2017-08-09]. https://arxiv.org/pdf/1708.02862.pdf

Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.324]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnik C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu Y, Zhang D S, Lu G J and Ma W Y. 2007. A survey of content-based image retrieval with high-level semantics. Pattern Recognition, 40(1): 262-282 [DOI: 10.1016/j.patcog.2006.04.045]

Liu Y F, Cao J J, Li B, Yuan C F, Hu W M, Li Y X and Duan Y Q. 2019a. Knowledge distillation via instance relationship graph//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7096-7104 [DOI: 10.1109/CVPR.2019.00726]

Liu Y F, Chen K, Liu C, Qin Z C, Luo Z B and Wang J D. 2019b. Structured knowledge distillation for semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2604-2613 [DOI: 10.1109/CVPR.2019.00271]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Luo P, Zhu Z Y, Liu Z W, Wang X G and Tang X O. 2016. Face model compression by distilling knowledge from neurons//Proceedings of the 13th AAAI Conference on Artificial Intelligence. Phoenix, USA: AAAI: 3560-3566 [DOI: 10.5555/3016387.3016404]

Malinin A, Mlodozeniec B and Gales M. 2019. Ensemble distribution distillation [EB/OL]. [2019-11-25]. https://arxiv.org/pdf/1905.00076.pdf

Mehta R and Ozturk C. 2018. Object detection at 200 frames per second//Proceedings of 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 659-675 [DOI: 10.1007/978-3-030-11021-5_41]

Mirza M and Osindero S. 2014. Conditional generative adversarial nets [EB/OL]. [2014-11-06]. https://arxiv.org/pdf/1411.1784.pdf

Mirzadeh S I, Farajtabar M, Li A, Levine N, Matsukawa A and Ghasemzadeh H. 2020. Improved knowledge distillation via teacher assistant//Proceedings of 2020 AAAI Conference on Artificial Intelligence, 34(4): 5191-5198 [DOI: 10.1609/aaai.v34i04.5963]

Mishra A and Marr D. 2017. Apprentice: using knowledge distillation techniques to improve low-precision network accuracy [EB/OL]. [2017-11-15]. https://arxiv.org/pdf/1711.05852.pdf

Park J S, Li S, Wen W, Tang P T P, Li H, Chen Y R and Dubey P. 2017. Faster CNNs with direct sparse convolutions and guided pruning [EB/OL]. [2017-07-28]. https://arxiv.org/pdf/1608.01409.pdf

Park S Y and Heo Y S. 2020. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. Sensors, 20(16): #4616 [DOI: 10.3390/s20164616]

Park W P, Kim D J, Lu Y and Cho M S. 2019. Relational knowledge distillation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3967-3976 [DOI: 10.1109/CVPR.2019.00409]

Passalis N, Tzelepi M and Tefas A. 2020. Heterogeneous knowledge distillation using information flow modeling//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE: 2339-2348 [DOI: 10.1109/CVPR42600.2020.00241]

Peng B Y, Jin X, Li D S, Zhou S F, Wu Y C, Liu J H, Zhang Z N and Liu Y. 2019. Correlation congruence for knowledge distillation//Proceedings of 2019 IEEE International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5007-5016 [DOI: 10.1109/ICCV.2019.00511]

Phuong M and Lampert C H. 2019. Distillation-based training for multi-exit architectures//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1355-1364 [DOI: 10.1109/ICCV.2019.00144]

Polino A, Pascanu R and Alistarh D. 2018. Model compression via distillation and quantization [EB/OL]. [2018-02-15]. https://arxiv.org/pdf/1802.05668.pdf

Radosavovic I, Dollár P, Girshick R, Gkioxari G and He K M. 2018. Data distillation: towards omni-supervised learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4119-4128 [DOI: 10.1109/CVPR.2018.00433]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788 [DOI: 10.1109/CVPR.2016.91]

Ren S Q, He K M, Girshick R and Sun J. 2017. Faster R-CNN: towards

real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149 [DOI: 10.1109/TPAMI.2016.2577031]

Romero A, Ballas N, Kahou S E, Chassang A, Gatta C and Bengio Y. 2015. Fitnets: hints for thin deep nets [EB/OL]. [2015-03-27]. https://arxiv.org/pdf/1412.6550.pdf

Ruder S, Ghaffari P and Breslin J G. 2017. Knowledge adaptation: teaching to adapt [EB/OL]. [2017-02-07]. https://arxiv.org/pdf/1702.02052.pdf

Sainath T N, Kingsbury B, Sindhwani V, Arisoy E and Ramabhadran B. 2013. Low-rank matrix factorization for deep neural network training with high-dimensional output targets//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE: 6655-6659 [DOI: 10.1109/ICASSP.2013.6638949]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2019. MobileNetV2: inverted residuals and linear bottlenecks [EB/OL]. [2019-03-21]. https://arxiv.org/pdf/1801.04381.pdf

Sau B B and Balasubramanian V N. 2016. Deep model compression: distilling knowledge from noisy teachers [EB/OL]. [2016-11-02]. https://arxiv.org/pdf/1610.09650.pdf

Shao S, Li Z M, Zhang T Y, Peng C, Yu G, Zhang X Y, Li J and Sun J. 2019. Objects365: a large-scale, high-quality dataset for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8430-8439 [DOI: 10.1109/ICCV.2019.00852]

Shapiro L. 2019. Computer vision introduction [EB/OL]. [2019-02-22]. https://courses. cs. washington. edu/courses/cse473/19wi/notes/Vision1-19.pdf

Shu C Y, Liu Y F, Gao J F, Yan Z and Shen C H. 2021. Channel-wise knowledge distillation for dense prediction [EB/OL]. [2021-01-22]. https://arxiv.org/pdf/2011.13256.pdf

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2015-04-10]. https://arxiv.org/pdf/1409.1556.pdf

Sinha R K, Pandey R and Pattnaik R. 2018. Deep learning for computer vision tasks: a review [EB/OL]. [2018-04-11]. https://arxiv.org/ftp/arxiv/papers/1804/1804.03928.pdf

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Tian Y L, Krishnan D and Isola P. 2022. Contrastive representation distillation [EB/OL]. [2022-01-24]. https://arxiv.org/pdf/1910.10699.pdf

Tung F and Mori G. 2018. Deep neural network compression by in-parallel pruning-quantization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(3): 568-579 [DOI: 10.1109/TPAMI.2018.2886192]

Tung F and Mori G. 2019. Similarity-preserving knowledge distillation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 1365-1374 [DOI: 10.1109/ICCV.2019.00145]

Voulodimos A, Doulamis N, Doulamis A and Protopapadakis E. 2018. Deep learning for computer vision: a brief review. Computational Intelligence and Neuroscience, 2018: #7068349 [DOI: 10.1155/2018/7068349]

Wang D D, Li Y D, Wang L Q and Gong B Q. 2020a. Neural networks are more productive teachers than human raters: active Mixup for data-efficient knowledge distillation from a Blackbox model//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE: 1498-1507 [DOI: 10.1109/CVPR42600.2020.00157]

Wang T, Yuan L, Zhang X P and Feng J S. 2019. Distilling object detectors with fine-grained feature imitation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4933-4942 [DOI: 10.1109/CVPR.2019.00507]

Wang X L, Kong T, Shen C H, Jiang Y N and Li L. 2020b. SOLO: segmenting objects by locations [EB/OL]. [2020-07-19]. https://arxiv.org/pdf/1912.04488.pdf

Wang Y K, Zhou W, Jiang T, Bai X and Xu Y C. 2020c. Intra-class feature variation distillation for semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Virtual: Springer: 346-362 [DOI: 10.1007/978-3-030-58571-6_21]

Wang Z Y, Deng Z D and Wang S Y. 2016. Accelerating convolutional neural networks with dominant convolutional kernel and knowledge pre-regression//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 533-548 [DOI: 10.1007/978-3-319-46484-8_32]

Wu J X, Leng C, Wang Y H, Hu Q H and Cheng J. 2016. Quantized convolutional neural networks for mobile devices//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4820-4828 [DOI: 10.1109/CVPR.2016.521]

Wu Y. 2007. An introduction to computer vision [EB/OL]. [2007-03-26]. http://users. eecs. northwestern. edu/~yingwu/teaching/EECS432/Notes/intro.pdf

Xie J F, Shuai B, Hu J F, Lin J Y and Zheng W S. 2018. Improving fast segmentation with teacher-student learning [EB/OL]. [2018-10-19]. https://arxiv.org/pdf/1810.08476.pdf

Xie S N, Girshick R, Dollár P, Tu Z W and He K M. 2017. Aggregated residual transformations for deep neural networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1492-1500 [DOI: 10.1109/CVPR.2017.634]

Xu Z, Hsu Y C and Huang J W. 2018. Training shallow and thin net-

works for acceleration via knowledge distillation with conditional adversarial networks [EB/OL]. [2018-04-16]. https://arxiv.org/pdf/1709.00513.pdf

Yang C L, Xie L X, Su C and Yuille A L. 2019. Snapshot distillation: Teacher-student optimization in one generation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2859-2868 [DOI: 10.1109/CVPR.2019.00297]

Yim J H, Joo D G, Bae J H and Kim J M. 2017. A gift from knowledge distillation: fast optimization, network minimization and transfer learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4133-4141 [DOI: 10.1109/CVPR.2017.754]

Yu F and Koltun V. 2016. Multi-scale context aggregation by dilated convolutions [EB/OL]. [2016-04-30]. https://arxiv.org/pdf/1511.07122.pdf

Yun S M, Park J J, Lee K M and Shin J W. 2020. Regularizing class-wise predictions via self-knowledge distillation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE: 13876-13885 [DOI: 10.1109/CVPR42600.2020.01389]

Zagoruyko S and Komodakis N. 2017a. Wide residual networks [EB/OL]. [2017-06-14]. http://arxiv.org/pdf/1605.07146.pdf

Zagoruyko S and Komodakis N. 2017b. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer [EB/OL]. [2017-02-12]. https://arxiv.org/pdf/1612.03928.pdf

Zeng Z Y, Liu B, Fu J L, Chao H Y and Zhang L. 2019. WSOD2: learning bottom-up and top-down objectness distillation for weakly-supervised object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8292-8300 [DOI: 10.1109/ICCV.2019.00838]

Zhang H, Wu C R, Zhang Z Y, Zhu Y, Lin H B, Zhang Z, Sun Y, He T, Mueller J, Manmatha R, Li M and Smola A. 2022. ResNeSt: split-attention networks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop. New Orleans, USA: IEEE: 2736-2746 [DOI: 10.1109/CVPRW56347.2022.00309]

Zhang H Y, Cisse M, Dauphin Y N and Lopez-Paz D. 2018. mixup: beyond empirical risk minimization [EB/OL]. [2018-04-27]. https://arxiv.org/pdf/1710.09412.pdf

Zhang L F and Ma K S. 2021. Improve object detection with feature-based knowledge distillation: towards accurate and efficient detectors [EB/OL]. [2021-03-16]. https://openreview.net/pdf?id=uKhGRvM8QNH

Zhang L F, Song J B, Gao A, Chen J W, Bao C L and Ma K S. 2019. Be your own teacher: improve the performance of convolutional neural networks via self distillation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3713-3722 [DOI: 10.1109/ICCV.2019.00381]

Zhang Z, Ning G H and He Z H. 2017. Knowledge projection for deep neural networks [EB/OL]. [2017-10-26]. https://arxiv.org/pdf/1710.09505.pdf

Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2881-2890 [DOI: 10.1109/CVPR.2017.660]

Zhou A J, Yao A B, Guo Y W, Xu L and Chen Y R. 2017. Incremental network quantization: towards lossless CNNs with low-precision weights [EB/OL]. [2017-08-25]. https://arxiv.org/pdf/1702.03044.pdf

Zhou X Y, Wang D Q and Krähenbühl P. 2019. Objects as points [EB/OL]. [2019-04-25]. https://arxiv.org/pdf/1904.07850.pdf

作者简介

孙若禹，男，硕士研究生，主要研究方向为计算机视觉、知识蒸馏和目标检测。E-mail：sunruoyu1@alumni.sjtu.edu.cn

熊红凯，通信作者，男，教授，主要研究方向为信号处理与编码、多媒体网络通信、数据表示与学习。

E-mail：xionghongkai@sjtu.edu.cn