文章编号:1672-3961(2025)04-0056-16 DOI:10.6040/j.issn.1672-3961.0.2024.055

基于生成对抗网络的知识蒸馏研究综述

杨巨成,路开奎,王媛
(天津科技大学人工智能学院, 天津 300457)

摘要:为了总结生成对抗网络在知识蒸馏中的应用,探索生成对抗网络在知识蒸馏中的协同机制与优化潜力,本研究开展基于生成对抗网络的知识蒸馏研究综述。概述基于输出层特征的知识蒸馏、基于中间层特征的知识蒸馏、基于关系特征的知识蒸馏、基于结构特征的知识蒸馏 4 种知识蒸馏研究进展,分析不同知识蒸馏形式的优势和缺点。详细介绍基于生成对抗网络的知识蒸馏分类和相关研究进展。针对生成对抗网络的知识蒸馏技术存在的局限性,给出未来优化方向和拓展应用方向。

关键词:知识迁移;模型压缩;知识蒸馏;模型轻量化;生成对抗网络
中图分类号:TP391 **文献标志码**:A
引用格式:杨巨成,路开奎,王媛. 基于生成对抗网络的知识蒸馏研究综述[J]. 山东大学学报(工学版),2025,55(4):56-71.
YANG Jucheng, LU Kaikui, WANG Yuan. Review of knowledge distillation based on generative adversarial networks[J]. Journal of Shandong University (Engineering Science), 2025, 55(4):56-71.

Review of knowledge distillation based on generative adversarial networks

YANG Jucheng, LU Kaikui, WANG Yuan
(College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China)

Abstract: To summarize the application of generative adversarial networks in knowledge distillation, and explore the collaborative mechanisms and optimization potential of generative adversarial networks in knowledge distillation, a review of knowledge distillation based on generative adversarial networks was conducted. Research progress was reviewed in four categories of knowledge distillation, including methods based on output features, intermediate features, relational features, and structural features. The advantages and disadvantages of each approach were analyzed. The classification and development of knowledge distillation methods based on generative adversarial networks were introduced in detail. Limitations of these knowledge distillation techniques based on generative adversarial networks were identified, and potential directions for optimization and application expansion were proposed.

Keywords: knowledge transfer; model compression; knowledge distillation; model lightweighting; generative adversarial networks

0 引言

目前,人工智能产业发展迅速,深度学习在各个研究领域展现出非凡性能,带动多学科共同发展。随着实际应用场景的复杂化,深度神经网络通过不断加深网络结构解决实际问题,使网络参数、计算量呈指数式增加,无法将其部署在算力受限的移动端设备。为了让高性能的深度神经网络部署在资源受限的低资源设备上,需要对复杂网络进行优化压缩,其中知识蒸馏(knowledge distillation, KD)是一种新型的网络模型压缩方法[1]。知识蒸馏涉及两个模型,通过将一个预训练的大规模复杂网络模型(教师网络)学到的知识迁移到另一个小规模轻量的网络模型(学生网络)上,实现模型轻量化。相较于其他模型压缩方法,知识蒸馏的优势在于可以将复杂网络压缩成任意结构的轻量化网络,将教师网络中的知识迁移到学生网络中,提高学生网络性能。目前,知识蒸馏已经演变出各式各样的变体和策略,逐渐从蒸馏输出层演变至蒸馏中间特征层[2-6]。随着对知识蒸馏的进一步研究,教师网络传递的知识变得更加多元化,从中间特征知识到

收稿日期:2024-03-12
基金项目:天津市研究生科研创新资助项目(2022SKYZ370)
第一作者简介:杨巨成(1980—),男,湖北天门人,教授,博士生导师,博士,主要研究方向为人工智能和智能驾驶。E-mail:jcyang@ tust.edu.cn

结构特征知识[7-9],从单一节点到多节点[10-12]。知识蒸馏是以获取有效的可迁移知识实现模型压缩,没有改变网络中的参数、通道、卷积核等。所以,知识蒸馏技术可以联合其他压缩和加速方法共同实现模型轻量化,如参数量化[13]、模型剪枝[14]、轻量化卷积和设计[15]等。知识蒸馏技术在各种任务和场景中展现出优势,例如:计算机视觉中的分类任务[16-17]、目标检测任务[18]、语义分割任务[19],自然语言处理中预训练模型的压缩,高性能推荐系统模型的压缩,等等。知识蒸馏在各领域发挥重要作用,对于知识蒸馏技术的研究具有一定的理论意义和广泛的应用价值。

文献[20]提出一种无监督深度学习范式——生成对抗网络(generative adversarial networks, GANs),通过生成器和判别器进行对抗训练,使生成器生成特定的数据特征分布空间。利用 GANs 可以生成各种高质量图片,广泛应用于图像生成[21]、图像分割[22]等领域,研究前景广阔。GANs 的生成器可以模拟特定的数据分布,与知识蒸馏的核心思想(即学生网络模拟教师网络输出分布)类似。因此,GANs 结合知识蒸馏技术成为新的研究方向,得到从事轻量化深度神经网络学者的广泛关注。研究发现,二者结合的优势主要体现在提高模型性能和效率方面。在 GANs 与知识蒸馏的协同框架中,生成器通过捕捉教师网络的目标数据分布特性,生成覆盖更广泛的合成样本。这一机制尤其适用于原始训练数据有限或分布不均衡的场景,为学生网络提供多样化的增强数据。即便面对生成样本潜在的不稳定性,该融合方法仍能引导学生网络从多维度数据分布中提炼鲁棒特征表示,显著提升模型的跨域泛化能力、小样本学习精度、任务鲁棒性等核心性能指标。在传统知识蒸馏过程中,学生网络直接从教师网络中学习。在基于 GANs 的知识蒸馏中,学生网络可以从生成器产生的高质量数据中学习,学习过程更加高效,生成的数据可以针对性地包含教师网络的关键知识,而不是简单地从教师网络输出中进行学习,减少学生网络训练所需的时间和资源。基于 GANs 的特性,使用 GANs 生成的数据可以减少对大量标记数据的需求,对难以获得大量标记数据的应用场景具有一定价值。知识蒸馏可以利用有限的数据更高效地训练模型,通过从教师网络中蒸馏知识,使小模型在数据稀缺的情况下取得良好性能,在零样本学习上能够发挥巨大潜力。

目前,关于知识蒸馏技术的综述重点以归纳知识的不同形式及对比各种关键方法为主[23-24],对知识蒸馏与交叉领域结合的描述较少。本研究聚焦基于 GANs 的知识蒸馏领域,对知识蒸馏相关原理进行详细阐述,对基于 GANs 的知识蒸馏技术和主要应用进行展开,分别从知识蒸馏原理、知识蒸馏的知识形式、基于 GANs 的知识蒸馏研究现状、问题与挑战、前景展望 5 个方面进行介绍,方便相关学者进一步探索知识蒸馏领域。

1　知识蒸馏理论基础

文献[25]提出一种模型压缩方法,使一个新模型近似地模拟一个预训练好的原模型,原模型为新模型提供额外的监督信息,实现模型压缩;文献[26]通过 L2 损失函数训练小型网络,使小型网络学习大型网络输出的逻辑单元,研究证明,从大型网络集合中获得的知识可以成功迁移到结构相对简单的小型网络中;文献[27]利用深度神经网络的输出分布特性,将小型网络和大型网络输出分布之间的 Kullback-Leibler(KL)散度最小化,实现模型压缩。然而,上述方法通常使用逻辑层或类别概率表示知识,存在噪声信息或信息丢失问题,限制学生网络的泛化能力。

为了解决上述问题,文献[1]提出知识蒸馏概念,首次提出大型教师网络的暗知识提取,强调知识迁移过程和知识类型之间的关系,通过试验证明知识蒸馏在深度神经网络中的有效性。知识蒸馏是将大型教师网络的知识传递到小型学生网络中。这个传递过程通过学习大型网络 Softmax 函数的软化类别分布输出实现,学生网络通过该过程学习教师网络的泛化能力,因此,经过知识蒸馏训练的学生网络性能通常优于仅拟合训练数据的学生网络。知识蒸馏原理框架如图 1 所示,其中 T 为温度系数,用于控制输出概率的软化程度;t 为实际温度系数。

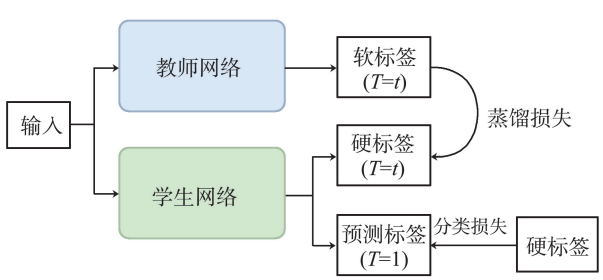

图 1　知识蒸馏原理框架

Fig.1　Knowledge distillation principle framework

文献[1]提出的知识蒸馏利用经过改进的 Softmax 层将逻辑层的输出概率(带有 T 的类概率)软化,软化的类概率作为传递知识诱导学生网络训

练，实现知识迁移。其中，将软化后的类概率称为软标签，原始数据的标签称为硬标签。软标签中不仅包含类别信息，也包含类别之间关系的隐含信息，和硬标签相比具有更多可学习的内容。在分类任务中，深度学习网络最后得到的不正确类别的分类概率可能非常小，但不正确类别之间的相对概率可以很好地表示模型如何得到更好的泛化能力。通过在输出层加入 T 得到软标签，则神经网络第 i 类概率输出

$$q_i=\frac{\exp(z_i/T)}{\sum_{i=1}^{j}\exp(z_i/T)},\tag{1}$$

式中，z_i 为神经网络第 i 类逻辑输出，$i\in\{1,2,\cdots,j\}$。

用软标签让学生网络学习不同类别间关系的隐含知识，得到更好的学习效果，提高泛化能力。软标签、硬标签、学生网络预测输出三者共同指导学生网络训练，学生网络的损失函数

$$L_{\mathrm{s}}=(1-\lambda)L_{\mathrm{CE}}(y_{\mathrm{true}},p_{\mathrm{s}})+\lambda L_{\mathrm{CE}}(q_{\mathrm{s}},q_{\mathrm{t}}),\tag{2}$$

式中：p_{s} 为学生网络的预测概率；y_{true} 为样本的真实标签；L_{CE} 为交叉熵损失函数；λ 为平衡参数，可以选择固定值，也可以选择动态调整；q_{s} 为学生网络软标签；q_{t} 为教师网络软标签。

在暗知识提取背景下[1]，知识蒸馏的主要工作是利用 T 输出软标签，实现标签平滑和置信度惩罚作用，为学生网络提供正则化约束[27-28]。标签平滑是在计算损失函数时，将正确标签的概率分配到其他类别，缓解过拟合，目的是让模型更加关注分类问题的结构，而不是过分相信训练样本的真实标签。置信度惩罚是在训练过程中，通过教师网络的置信度指导学生网络训练，提高模型的泛化能力，当学生网络在某个样本上的预测与教师网络的预测差异显著时，学生网络的输出概率分布将通过蒸馏损失函数受到惩罚。这种约束机制强制学生网络学习教师网络的决策边界特性，提升知识迁移效率与模型泛化能力。虽然标签平滑和置信度惩罚均提供正则化，但二者的实现方式和目的略有不同。标签平滑侧重于减少过拟合，置信度惩罚侧重于提高泛化能力。

2 知识蒸馏的知识形式

知识蒸馏属于迁移学习，文献[1]提出的知识蒸馏将已训练的模型作为教师网络，提取其中的软标签知识，用于训练学生网络，希望轻量级的学生网络能够学到与教师网络相似的性能。在设计学生网络时，需要建立学生网络的某些知识与教师网络的对应关系，可以是师生网络间特征图、注意力图、网络结构、样本信息等知识的关系，通过某些方法（如参数调整、参数共享）增强师生网络间该类型知识的相似性，完成模型压缩。知识蒸馏是一种简单有效的方法，能够从强大的教师网络中学习更多的知识，用于提高相对紧凑的学生网络性能。目前，在知识蒸馏研究与应用中，学生网络和教师网络处理的任务基本相同，即教师网络和学生网络的输出是一致的。而在不同的知识蒸馏策略中，教师网络输出和学生网络输出之间的关系可能有所不同，具体取决于采用的蒸馏方法、蒸馏知识类型及技术。根据迁移知识形式的不同，本研究将知识蒸馏分为以下4个类型：基于输出层特征的知识蒸馏、基于中间层特征的知识蒸馏、基于关系特征的知识蒸馏、基于结构特征的知识蒸馏。

2.1 基于输出层特征的知识蒸馏

基于输出层特征的知识通常指教师网络最后一个输出层的分布输出。用大规模深度神经网络的输出作为教师知识，让小规模深度神经网络直接模仿教师网络的最终预测。基于输出层特征的知识蒸馏是一种简单有效的模型压缩方法，广泛应用于不同任务和应用中，其通用结构如图2所示。

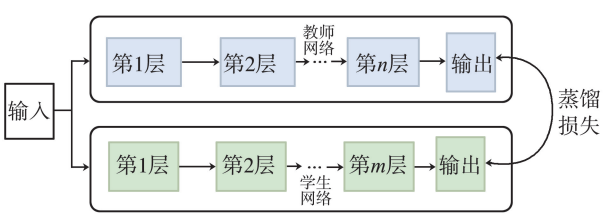

图2 基于输出层特征的知识蒸馏结构示意图

Fig.2 Schematic diagram of knowledge distillation structure based on output layer characteristics

文献[1]提出的知识蒸馏方法是最基础的基于输出层特征的知识蒸馏方法，其原理是控制最后一个输出层分布输出的软化程度传递知识。对于如何充分利用教师网络最后一个输出层的分布输出，研究学者提出除基础知识蒸馏以外的其他方法。为了避免学生网络学习不重要的输出分布，大多数研究内容通过改进置信度惩罚、标签正则化优化原始的知识蒸馏[29-32]。文献[29]提出最高分差（top score difference，TSD）知识蒸馏，创新性引入置信度校准机制，通过在教师网络的训练过程中施加动态正则化约束，优化预测分布的置信度校准特性；

通过调节教师网络对预测不确定性的表征能力,有效缓解传统方法中因教师网络过度自信导致的标签噪声问题。虽然该方法中教师网络的预测精度出现可控范围内的下降,但学生网络通过吸收教师网络校正后的知识表达,在基准数据集上展现出持续增强的泛化性能。当没有强大的教师网络可以应用时,文献[30]提出一种无需教师的知识蒸馏(teacher-free knowledge distillation, Tf-KD)框架,其中学生网络从自身学习或手动设计标签平滑实现正则化分布,输出结果不再是单一、硬性的标签,而是柔和的概率分布,使模型更加鲁棒,更容易泛化到新的数据。该方法通用性更强,可与大规模网络蒸馏后的结果相媲美。文献[31]将经典的 KD 损失转化为两部分,即目标类知识蒸馏(target class knowledge distillation, TCKD)和非目标类知识蒸馏(non-target class knowledge distillation, NCKD),通过实证研究发现,TCKD 转移有关训练样本难度的知识,NCKD 是使逻辑层蒸馏起作用的重要原因。经典 KD 抑制 NCKD 的有效性,限制平衡 TCKD 和 NCKD 的灵活性。为了解决这一问题,文献[31]提出解耦知识蒸馏(decoupled knowledge distillation, DKD),使 TCKD 和 NCKD 更有效、更灵活地发挥作用。

使用基于输出层特征的知识蒸馏最大优势在于无需考虑师生网络之间内在结构的相对差异,仅利用模型对样本的输出完成蒸馏,可以应用在跨模态[33]、自监督[34]、自蒸馏[35]、协同蒸馏[36]等领域,还可与关系型知识[37]、中间层知识[38]等其他蒸馏形式相结合,实现多种知识蒸馏。但该策略的局限性在于输出信息有限,无法模拟教师网络中间信息,仅适用于传统的分类任务,在较为高级的视觉任务上无法获取样本更多的上下文信息,不够灵活。

2.2　基于中间层特征的知识蒸馏

为了使学生网络从教师网络中获得更多知识,改进基于输出层特征的知识蒸馏策略,研究学者在知识的设计上进行创新。基于中间层特征的知识来源于网络模型特征提取层的输出,一般代表中间层提取的不同尺度的特征信息。通过减少师生网络间的中间层特征差异,学生网络模拟教师网络提取的中间特征满足高级任务对中间层知识的需求,为学生网络提供更多监督信号,解决基于输出特征知识中存在的信息单一问题。基于中间层特征的知识蒸馏结构如图 3 所示。基于中间层特征的知识蒸馏方法层出不穷,如特征图蒸馏、注意力图蒸馏、激活层蒸馏等[5,38-45]。与基于输出层特征的知识蒸馏不同,该方式需要考虑师生网络间的结构差异,可分为同构蒸馏和异构蒸馏。同构蒸馏是师生网络架构相似或属于同一网络结构,层与层之间一一对应,如残差网络(residual network, ResNet)系列、视觉几何组(visual geometry group, VGG)网络系列。异构蒸馏是师生网络中间层的特征图无法实现层与层之间的匹配,需要对学生特征做进一步处理才能继续蒸馏。本研究列举基于中间层特征的知识蒸馏代表性方法,详细分析各方法的优点、缺点及相关实现方式,如表 1 所示。

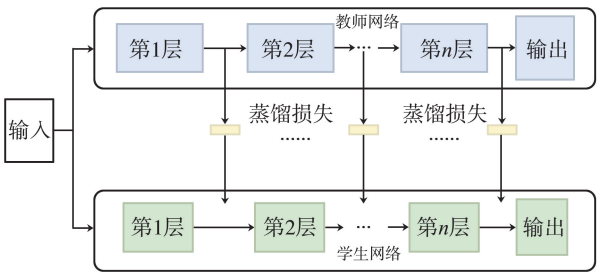

图 3　基于中间层特征的知识蒸馏结构示意图

Fig.3　Schematic diagram of knowledge distillation structure based on intermediate layer characteristics

深度神经网络具有多层次的特征信息可供学生网络学习,其中中间层知识代表特征提取层所提取的不同维度的特征知识。学生网络模仿学习教师网络的特征表达就是基于中间层知识蒸馏的核心思想。基于中间层特征的知识蒸馏学生网络损失函数

$$L_s(f_t(x),f_s(x))=L_F(\Phi_t(f_t(x)),\Phi_s(f_s(x))),\tag{3}$$

式中:$f_t(x)$和$f_s(x)$分别为教师和学生网络中间层的特征表达;$\Phi_t(\cdot)$和$\Phi_s(\cdot)$分别为教师和学生特征适配函数,主要作用是实现师生网络特征图维度对齐,常用于异构知识蒸馏中;L_F 为计算特征图之间差异性的函数。

当前,基于中间层特征的知识蒸馏方法十分丰富,不局限于传统的分类任务,在目标检测、实例分割等任务上也都表现俱佳,对特定场景下的特定任务可以找到相应策略完成网络压缩任务。但由于目前策略过于冗余,缺乏整合各种类型中间层特征知识的方法实现对不同角度、不同维度知识的传递。另外,如何统一同构蒸馏或异构蒸馏间中间层特征知识的传递,也是值得探讨的问题。

表1 基于中间层特征的知识蒸馏代表性方法

Table 1 Representative methods of knowledge distillation based on intermediate layer characteristics

类别	代表方法	蒸馏损失函数	优点	缺点	实现方式
同构蒸馏	AT[38]	$L_2(\cdot)$	首次提出转移注意力图知识,能与KD结合,在多个数据集实现性能提升	需要一些经验或基于试验调整超参数,用于平衡分类损失和注意力转移损失的权重	用空间注意力映射函数获得并转移中间层空间注意力图实现知识蒸馏,学生网络能够更好地关注图像中的关键区域,提高性能和泛化能力
	SP[39]	$L_2(\cdot)$	有效传递中间层知识和类间关系知识	对于样本数量的需求和计算成本较高	对比师生网络之间成对数据的相似度矩阵差异,优化目标函数
	PKT[40]	$L_{CE}(\cdot)$	有效利用特征空间中数据的概率分布传递中间层特征知识	数据敏感性高,当数据特征分布稀疏时,存在信息丢失现象,无法得到预期的性能提升	匹配师生数据在特征空间中的概率分布进行知识蒸馏,将教师网络的中间特征空间结构映射到学生中间特征空间中
	DFA[41]	$L_2(\cdot)$	利用二阶可微的特征聚合方法(即可用单教师网络模拟多教师蒸馏),减少计算资源	过度依赖同构蒸馏,使用范围过于局限	将教师网络的多层特征聚合,通过网络结构搜索进行加权聚合,实现中间层特征知识转移
	KDSVD[42]	$L_2(\cdot)$	压缩知识信息,减少内存和计算量	需要对模型的每个训练步骤执行奇异值分解,导致训练时间增加	用奇异值分解去除中间层特征中的空间冗余信息,提取核心知识,传递隐含特征信息
异构蒸馏	FitNet[5]	$L_2(\cdot)$	拓宽知识定义,用中间层解决网络压缩	需手动选定某一层间的知识,不灵活	适配师生网络间的中间特征,使用均方误差衡量两者差异
	AB[43]	$L_2(\cdot)$	关注激活边界在网络中的重要性,证明激活传递损失在知识传递中比均方误差更有效	由于增加额外的梯度下降过程,时间开销大	通过让教师网络与学生网络之间的激活边界(ReLU函数知识信息)尽可能一致,实现知识转移
	FT[44]	$L_2(\cdot)$	通过卷积操作提出共享特征空间概念,使教师网络的中间层知识易于学生网络接受	需要以无监督的方式额外训练解释器	用卷积操作解释教师网络知识并解码给学生网络,使学生网络内化为自己的知识
	VID[45]	$L_1(\cdot)$	证明卷积神经网络和多层感知机之间异构知识转移的可能性	需要额外的卷积层和激活层,计算开销较高	师生网络在特征提取过程中保持高互信息,通过学习并估计教师网络中的分布,激发知识传递,使相互信息最大化

注:注意力转移(attention transfer, AT)、相似性保留(similarity-preserving, SP)、概率知识转移(probabilistic knowledge transfer, PKT)、可微特征聚合(differentiable feature aggregation, DFA)、奇异值分解知识蒸馏(knowledge distillation using singular value decomposition, KDSVD)、轻量深度网络的知识引导(hints for thin deep nets, FitNets)、激活边界(activation boundaries, AB)、因子转移(factor transfer, FT)、变分信息蒸馏(variational information distillation, VID)。$L_2(\cdot)$为均方误差损失函数,$L_1(\cdot)$为平均绝对误差损失函数。

2.3 基于关系特征的知识蒸馏

基于关系特征的知识主要为教师网络中的关系知识。关系特征是指教师网络在不同层次和不同数据样本之间的关系知识,学生网络学习的核心不是点对点特征的输出结果,而是不同层次和样本数据之间的关系。该类型知识关键在于提供一个恒等的关系映射,帮助学生网络更有效地学习教师网络的关系特征知识。基于关系特征的知识可以根据需求自由构建不同的关系对象,包括不同层次之间的关系、不同样本之间的关系,甚至是层次和样本之间的多重关系。基于关系特征的知识蒸馏结构如图4所示,T_i、S_i分别为教师网络、学生网络提取到的不同层次的关系特征($i=1,2,\cdots,n$),$\omega(\cdot)$为构建关系特征知识的函数,R_t和R_s分别为教师网络、学生网络不同层次、不同样本之间的关系知识。以文本相关性任务为例,学生网络可以学习教师网络输出的不同文本样本对之间的相关性。

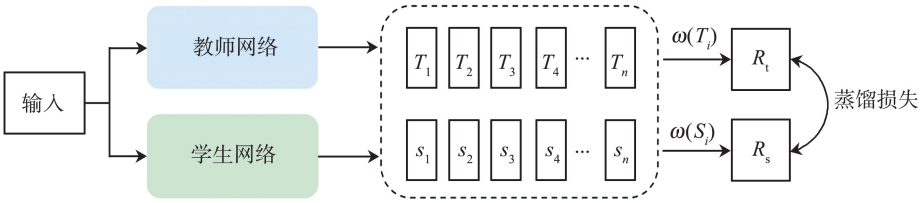

图 4　基于关系特征的知识蒸馏结构示意图

Fig.4　Schematic diagram of knowledge distillation structure based on relational characteristics

深度神经网络具有清晰的层次结构,从输入样本到输出概率是一个求解映射关系的过程,包含层与层之间的关系特征知识。文献[46]最早利用关系特征知识提出解决方案信息流(flow of solution procedure, FSP)蒸馏策略,将网络的中间特征转化为 FSP 矩阵,描述网络关系特征知识,让学生网络模仿生成与教师网络相似的 FSP 矩阵,实现模型压缩。具体实现方式如图 5 所示。

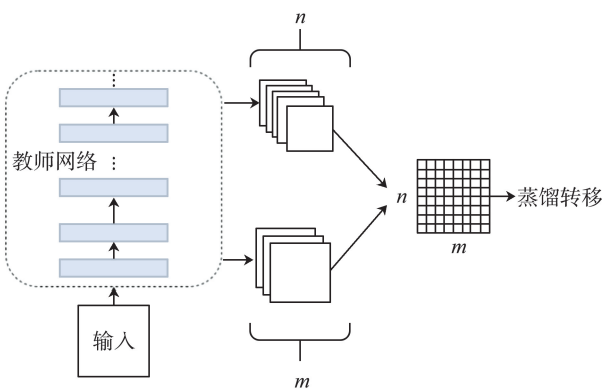

图 5　FSP 矩阵的关系特征知识迁移示意图

Fig.5　Relational characteristics knowledge transfer diagram of FSP matrix

FSP 蒸馏策略从基于中间层特征图关系出发,通过 FSP 矩阵差异传递关系特征知识。受此启发,通过近似师生网络中间层映射的雅可比矩阵[47]和径向基函数[42]也可以获取特征之间的相关性知识。文献[48]关注样本间的关系,提出样本间角度关系蒸馏和距离关系蒸馏,充分挖掘不同类别之间和相同类别之间的样本关系。一般基于样本关系的知识蒸馏传递的是样本之间的内在和外在关系信息,使学生网络形成与教师网络相同的样本关系,达到模型压缩的目的,如图 6 所示。基于关系特征的知识蒸馏策略大多不受师生网络架构之间的约束,可以通过其他辅助技术[如图知识蒸馏(graph knowledge distillation, GKD)[49]]获取关系特征知识完成知识迁移,有更高的灵活性。

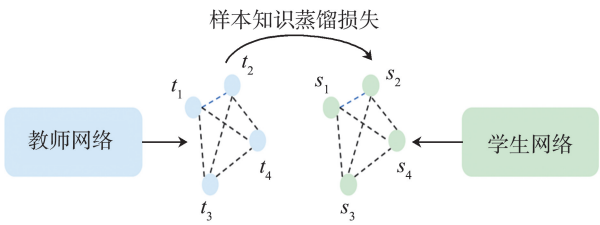

图 6　基于样本关系特征的知识迁移示意图

Fig.6　Knowledge transfer diagram based on sample relation characteristics

2.4　基于结构特征的知识蒸馏

深度学习网络性能不仅取决于网络参数和关系,还依赖于体系结构。基于结构特征的知识是教师网络的全面知识框架,从教师网络获取知识的结构体系出发,包括输出特征知识、中间特征知识、关系特征知识及教师网络的局部和全局特征分布。基于结构特征的知识蒸馏是一种整合其他知识传递方式的补充方法,旨在通过综合多种知识传递形式,促使学生网络的预测包含与教师网络相似的结构知识,使学生网络拥有无限接近甚至超过教师网络的性能。结构特征知识的构成因学生网络实现的下游任务不同而异,例如:为了优化原始 KD,利用样本特征、样本间关系、特征空间变换构成知识变化的结构知识[50];在多人姿态估计中,由输出特征、中间特征和全局预测特征组成整体结构知识[51];在道路场景分割任务中,将样本信息分割成不同区域作为节点,根据节点特征分布的相似性建立节点之间的成对关系,形成区域间亲和度的结构信息图知识[52]。

一般,基于结构特征的知识能够更加彻底和充分地利用教师网络。随着更多高级的计算机视觉、自然语言任务出现,更需要对结构特征知识进行迁移和提取。目前,基于结构特征的知识蒸馏还处于基础发展阶段,对于结构特征知识的挖掘大多借助其他技术获得,其中最为典型的是与 GANs 结合[53-54]。

2.5　性能比较

知识蒸馏通过对教师网络中知识的捕获和师生学习中的蒸馏策略,提升轻量级学生网络的性能

表现。为了更直观地体现知识蒸馏的有效性,本研究总结4种知识形式中典型的知识蒸馏方法在CIFAR100图像分类数据集上的分类性能,如表2所示,其中每种方法的试验分类准确率结果均来源于相应的原始文献。CIFAR100数据集由100个类别的32像素×32像素RGB(red, green, blue)图像组成,有50 000张训练图像和10 000张测试图像,每个类别都有相同数量的训练图像和测试图像。表2中由于Tf-KD方法是无需教师网络的自蒸馏方法,因此没有教师网络的表现性能。

表2 不同知识蒸馏方法在CIFAR100上的性能比较

Table 2 Performance comparison of different knowledge distillation methods on CIFAR100 单位:%

知识蒸馏类型	方法	教师网络	教师网络识别准确率	学生网络	学生网络识别准确率	知识蒸馏后学生网络识别准确率
基于输出层特征的知识蒸馏	KD[1]	ResNet56	72.34	ResNet20	69.06	70.66
	TSD[29]	DenseNet-190	82.83	DenseNet-100	77.20	82.75
	Tf-KD[30]	—	—	ResNet18	75.87	77.10
	DKD[31]	ResNet56	72.34	ResNet20	69.06	71.97
基于中间层特征的知识蒸馏	AT[38]	WRN-28-4	79.17	WRN-16-4	77.24	77.49
	SP[39]	WRN-28-4	79.17	WRN-16-2	73.42	74.09
	DFA[41]	WRN-28-4	79.17	WRN-16-4	77.24	79.74
	KDSVD[42]	VGG-T-DNN	64.44	VGG-S-DNN	61.37	65.05
	FT[44]	ResNet110	73.09	ResNet56	71.96	74.48
基于关系特征的知识蒸馏	FSP[46]	ResNet32	64.06	ResNet14	58.65	63.33
	RKD[48]	ResNet50	77.76	VGG11	71.26	74.66
	GKD[49]	WRN-28-1	68.74	WRN-28-0.5	61.50	61.83
基于结构特征的知识蒸馏	KDCAN[54]	WRN-40-10	79.38	WRN-10-4	71.48	74.25

注:“—”表示没有教师网络及对应的识别准确率。关系知识蒸馏(relational knowledge distillation, RKD)、结合条件对抗网络的知识蒸馏(knowledge distillation with conditional adversarial networks, KDCAN)。

3 基于GANs的知识蒸馏研究进展

在知识蒸馏的研究和应用中,解决教师网络和学生网络之间的全局一致性是一个重要研究内容。全局一致性关注的是师生网络在整个输入空间上行为的一致性,不仅仅局限于最终输出层的一致性,还包括网络中间层表示的一致性,即教师网络获得某类型知识的全过程,学生网络也可以完整复现。研究人员发现,在获取结构知识过程中,将知识蒸馏与某些拥有特殊机制的算法相结合也能实现知识传递。传统的蒸馏仅仅是传递单样本的输出信息,当教师网络很难在真实数据中学到知识时,学生网络也无法获得更好的性能。为解决该问题,研究人员对GANs产生极大兴趣。GANs的本质特性是通过对抗性学习生成新的分布,即生成器不断学习生成无法被判别器网络区分的分布,一般用于图像生成领域。GANs的主要结构包括生成器和判别器,生成器用于生成无限接近真实数据的样本分布,使判别器无法辨别。知识蒸馏是让学生网络模拟教师网络的相关分布信息,达到自身最优解空间,不断近似教师网络的性能。因此,将GANs的结构直接用于知识蒸馏过程,其中一方(通常是生成器作为学生网络)尝试生成模仿教师网络输出或特征的表示,另一方(判别器)尝试区分学生和教师网络的输出或特征。利用GANs的思想使教师网络和学生网络的输出知识趋于一致,实现网络压缩的目的[54]。利用GANs的知识蒸馏能够有效保持师生网络输出分布的多模态性质[55],通过相关方法可以实现向学生网络转移结构化知识体系,包括类间相关性、输出相关性、网络相关性等,大幅度提升学生网络的训练性能。

根据学生网络和教师网络担任身份的不同,基于GANs的知识蒸馏主要分为3类:学生网络担任生成器,利用判别器近似教师网络输出分布(即对生成器蒸馏);教师网络和学生网络担任判别器,指导生成器训练(即对判别器蒸馏);学生网络和教师网络共同担任生成器和判别器(即对生成器和判别器进行蒸馏)。

3.1 对生成器蒸馏

对生成器蒸馏是知识蒸馏和GANs相结合的早期成果。该方法把学生网络作为待训练生成器,教师网络作为预训练生成器,引入GANs的判别器与两种生成器做对抗训练。通常,该类型方法不仅在

输出层上从教师网络到学生网络进行知识传递,还在特征层面利用教师网络的内部特征表示指导生成器生成能够引导这些高级特征表示的数据,实现学生网络获取教师网络结构知识的目的。

3.1.1　对生成器蒸馏的实现过程

对生成器蒸馏需要定义一个蒸馏目标,确定生成器应如何受益于教师网络知识。该目标可以是生成特定类别的样本,也可以是生成能够最大化学生网络学习效率的样本。为实现这一目标,损失函数需要包含多个组成部分,一部分确保生成数据贴近真实数据分布,另一部分确保生成数据能够有效反映教师网络知识,例如将特征匹配损失与传统对抗损失结合使用,以鼓励生成器产生能够模仿教师网络特征表示的数据[54]。对生成器蒸馏的方法在训练过程中可以不需要标签,且能够推广到不同的师生网络中,主要思想是利用判别器区分学生网络产生的输出和预训练教师网络的输出结果,通过对抗训练在判别器中拟合师生网络产生的输出分布,判别器的更新是为了更好区分教师网络和学生网络的输出分布,学生网络的更新是为了更好地欺骗判别器,经过参数迭代更新训练,最终达到判别器无法区分师生网络输出的目的。

3.1.2　对生成器蒸馏的实际应用

对生成器蒸馏的方法能够提升学生网络担任的生成器生成高质量数据的能力,增强整个知识蒸馏流程的效果,广泛应用于图像生成领域。在该方法中,教师网络不仅指导学生网络学习目标任务,还指导生成器产生能够有效提升学生网络学习的数据。这意味着生成器不仅要学习数据分布,还要学习如何生成对学生网络最有用的数据,加强生成器与判别器之间的对抗。对生成器蒸馏的工作原理如图 7 所示。

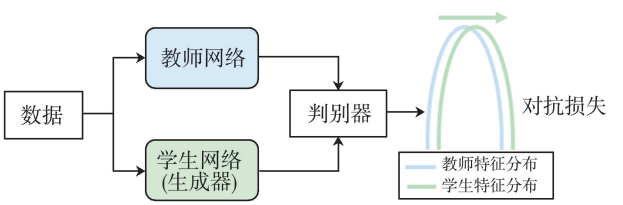

图 7　对生成器蒸馏示意图

Fig.7　Diagram of distillation against generator

文献[56]使用教师助理(判别器)协助学生网络学习教师网络知识,工作原理是将生成器网络看作一个具有非常少权值参数的学生神经网络,将判别器网络当作一个教师助理,用于区分学生神经网络和教师神经网络生成的特征,通过同时优化生成器网络和判别器网络,学生网络可以对输入数据生成具有教师网络同样的分布特征。该方法的损失函数

$$L_{\mathrm{GAN}}=\frac{1}{n}\sum_{i=1}^{n}\left[L_{\mathrm{CE}}(o_{\mathrm{s}}^{i},y^{i})+\delta L_{\mathrm{CE}}(\tau(o_{\mathrm{s}}^{i}),\tau(o_{\mathrm{t}}^{i}))\right]+\gamma\frac{1}{n}\sum_{i=1}^{n}\left[\ln(D(z_{\mathrm{t}}^{i}))+\ln(1-D(z_{\mathrm{s}}^{i}))\right],\quad(4)$$

式中,n 为样本数,o_{s}^{i} 为学生网络输出分布,o_{t}^{i} 为教师网络输出分布,y^{i} 为真实标签,$\tau(\cdot)$ 为输出软标签函数,z_{t}^{i} 和 z_{s}^{i} 分别为教师网络和学生网络经过生成器输出的分布,$D(\cdot)$ 为判别器输出结果,δ 和 γ 为平衡超参数。式(4)中,第一项为经典知识蒸馏损失,第二项为对抗损失。

文献[57]将教师网络和学生网络的中间特征知识作为真假样本输入判别器,使师生网络中间特征通过对抗性训练趋于相似,不需要标签信息监督,有更高的可拓展性。文献[58]中生成器是学生网络和教师网络,没有真正意义上的判别器,通过深度卷积生成对抗网络(deep convolutional generative adversarial network, DCGAN)预训练得到教师网络,交叉熵损失和二元交叉熵损失构成联合损失,优化生成器输出。文献[59]利用知识蒸馏生成对抗网络(knowledge distillation with generative adversarial networks, KDGAN)解决数据受限问题,该网络框架由学生网络担任的分类器、教师网络和鉴别器组成,通过优化师生网络间蒸馏损失和对抗性损失,学生网络能够学习真实数据分布。

上述工作在压缩 GANs 方面取得显著成果,但仍然存在潜在的模型冗余。文献[60]提出一种在线多粒度蒸馏方法对生成器蒸馏,获得轻量级 GANs,采用不同结构作为基础教师生成器,捕捉更多的互补知识,除输出层知识外,还结合师生生成器之间的中间层知识,减少 GANs 的计算量,同时获得更高的图像质量。为了进一步降低 GANs 的存储和计算需求,提高特征图的提取水平,加强压缩学生生成器生成高质量的图像,文献[61]提出判别器合作蒸馏(discriminator-cooperated distillation, DCD)方法,驱动学生生成器的中间结果学习教师生成器的相应输出,为避免出现 GANs 训练过程中的模式崩溃现象,构建一种协作对抗训练方法,即教师生成器与学生生成器共同训练。文献[62]提出类感知蒸馏方法对大规模条件生成对抗网络(conditional generative adversarial network, cGAN)进行压缩,利用教师网络和学生网络中间层注意力图的差异进行大规模知识转移。

3.2 对判别器蒸馏

随着知识蒸馏与 GANs 的结合,对判别器蒸馏逐渐受到关注。将学生网络视为待训练的判别器,教师网络作为预训练的判别器,通过某种方式将教师网络的知识传递给 GANs 中的判别器(学生网络)。对判别器蒸馏的核心在于将教师判别器模型的知识融入学生判别器模型,提升学生判别器区分真实样本与生成样本的能力,帮助学生判别器学习更复杂和细微的数据特征,指导生成器产生高质量的样本。该方法的优势在于生成器可以生成更难以被判别的样本欺骗判别器,与判别器进行对抗,提高整个 GANs 系统的性能。因此,该方法主要用于零样本或样本受限的学习领域,将生成器合成的数据用作训练集[63-65]。对判别器蒸馏示意图如图 8 所示。

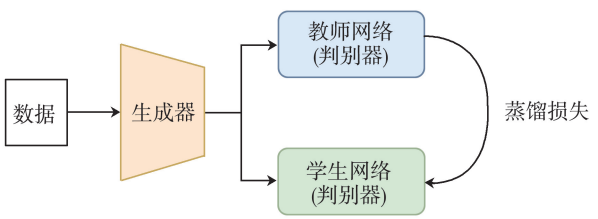

图 8 对判别器蒸馏示意图

Fig.8 Diagram of distillation against discriminator

3.2.1 对判别器蒸馏的实现过程

对判别器蒸馏的实现过程中,训练生成器产生合成数据样本,激活教师网络中的特定特征或类别响应。该方法通常包括:优化生成器,以最大化教师网络输出层的某些目标函数(如类别概率或特定特征图的激活函数);利用教师网络的输出分布和中间层的特定特征响应引导生成器生成能够模拟教师网络在真实数据上行为的合成数据,将学生判别器的训练目标调整为不仅区分真实数据和生成数据,还能从生成的合成数据中学习教师网络的行为,包括分类准确性和特征相似性。对判别器蒸馏在不需要原始训练数据的情况下,允许从教师网络向判别器传递知识。

3.2.2 对判别器蒸馏的实际应用

由于对判别器蒸馏方法的作用和优势,该方法广泛应用于零样本学习或样本受限领域中,也被称为零样本知识蒸馏(data-free knowledge distillation, DFKD)。为了克服由于隐私、安全性和保密性而无法获得预训练教师网络样本的问题,DFKD 中学生网络数据需要通过对抗学习策略进行合成。文献[63]提出零样本学习(data-free learning, DAFL)模型,将预先训练好的教师网络视为一个固定的判别器网络,生成器生成的训练样本可以在师生网络上传递最大响应的结构知识,使用生成数据和教师网络训练模型规模较小、计算复杂度较低的判别器学生网络,利用教师网络的 3 种先验信息(激活层信息、输出层信息、样本信息)作为结构知识进行传递;文献[66]通过训练一个对抗生成器搜索学生网络与教师网络匹配不佳的图片,用于训练学生网络,使零样本训练的学生网络能够和教师网络的预测相匹配。一般,DFKD 的损失函数

$$L_{\mathrm{DFKD}}=L_{\mathrm{G}}(F_{\mathrm{t}}(G(\boldsymbol{z})),F_{\mathrm{s}}(G(\boldsymbol{z}))), \tag{5}$$

式中,$F_{\mathrm{t}}(\cdot)$ 为教师网络输出,$F_{\mathrm{s}}(\cdot)$ 为学生网络输出,$G(\boldsymbol{z})$ 为随机向量 $\boldsymbol{z}$ 通过生成器生成的样本,$L_{\mathrm{G}}(\cdot)$ 为师生判别器之间的蒸馏损失函数。

文献[67]在 DAFL 基础上进行改进,提出零样本对抗蒸馏(data-free adversarial distillation, DFAD),在真实标签约束下,定义一个优化模型差异上界,使用 L_1 范数作为度量差异的指标,使 DFAD 更加高效和通用。与 DFAD 不同,文献[66]提出零概率知识传递(zero-shot knowledge transfer, ZSKT),利用 KL 散度衡量师生网络之间预测结果的度量差异,损失函数为

$$L_{\mathrm{S}}=D_{\mathrm{KL}}\left(T(\boldsymbol{x}_{\mathrm{p}})\,\|\,S(\boldsymbol{x}_{\mathrm{p}})\right)+\beta\sum_{l}^{N_{\mathrm{L}}}\left\|\frac{f(\boldsymbol{A}_l^{(\mathrm{t})})}{\|f(\boldsymbol{A}_l^{(\mathrm{t})})\|_2}-\frac{f(\boldsymbol{A}_l^{(\mathrm{s})})}{\|f(\boldsymbol{A}_l^{(\mathrm{s})})\|_2}\right\|_2, \tag{6}$$

式中:D_{KL} 为相对熵散度,用于衡量两个概率分布之间的差异;$T(\cdot)$ 和 $S(\cdot)$ 分别为预训练教师网络和学生网络判别器函数;$\boldsymbol{x}_{\mathrm{p}}$ 为生成器生成的伪造样本;$\boldsymbol{A}_l^{(\mathrm{t})}$ 和 $\boldsymbol{A}_l^{(\mathrm{s})}$ 分别为第 l 层教师网络和学生网络的激活块;N_{L} 为神经网络层数;$f(\cdot)$ 为归一化函数;β 为平衡参数。

文献[68]提出一种基于条件生成的零样本知识蒸馏(conditional generative data-free knowledge distillation, CGDD)框架,将预设标签视为真实标签训练半监督条件生成器,学生判别器通过注意力转移机制提取教师判别器中的隐藏知识,为学生网络提供更大的数据空间,提高蒸馏性能。

虽然以上几种方法都实现零样本知识蒸馏,但随着生成器更新,生成器和学生判别器进行额外的对抗性训练,使合成数据的分布发生变化,如果这种分布变化较大,则会使学生判别器遗忘前面步骤中获得的知识。为了缓解该问题,文献[69]提出动量对抗蒸馏(momentum adversarial distillation, MAD)方法,对生成器进行复制,利用复制生成器合成的样本进行训练,帮助学生判别器追溯已获得的知识,防止学生判别器过快适应生成器的更新,克

服合成数据分布变化过大的问题。在图像分类任务中,以上几种零样本知识蒸馏方法训练得到的学生网络在 CIFAR10 和 CIFAR100 数据集中的性能对比如表 3 所示。

表 3　不同零样本知识蒸馏方法训练的学生网络分类准确率

Table 3　Classification accuracy of the studentnetwork trained by different DFKD methods　　单位:%

数据集	教师网络	教师网络分类准确率	学生网络	学生网络分类准确率	使用零样本知识蒸馏方法后的学生网络分类准确率				
					DAFL[63]	DFAD[67]	ZSKT[66]	MAD[69]	CGDD[68]
CIFAR10	ResNet34	95.70	ResNet18	95.20	92.22	93.30	93.32	94.90	99.63
	WRN40-2	94.87	WRN16-2	93.95	81.55	—	89.66	92.64	—
CIFAR100	ResNet34	78.05	ResNet18	77.10	74.47	69.43	67.74	77.31	99.07
	WRN40-2	75.83	WRN16-2	73.56	40.00	—	28.44	64.05	—

注:“—”代表没有该类型方法对应的识别准确率。

随着对判别器蒸馏的进一步研究,文献[64]提出融合教师判别器知识策略,忽略学生网络,训练组栈式 GANs,利用多个教师网络作为鉴别器,重建与原始数据集近似的图像;文献[70]采用量化思想,对生成数据和教师判别器模型中的原始数据进行批量归一化层匹配统计,最大限度减少教师网络和经过量化的学生网络之间的差异;考虑当训练数据有限时,GANs 鉴别器出现严重的过拟合,文献[71]以对比语言-图像预训练(contrastive language-image pre-training, CLIP)模型作为教师网络提取知识,设计聚合生成知识蒸馏方法和相关生成知识蒸馏方法,减轻过拟合,有效提高生成器的生成性能和判别器的判别性能,弥补样本受限的劣势,提高生成样本的多样性。

对于难以获得原始数据的隐私性较高的方向(如生物特征识别),利用生成对抗蒸馏实现零样本和样本受限的知识蒸馏有巨大潜力。但教师网络的选择对蒸馏效果至关重要,理想的教师网络应该具有较好的性能和强大的特征表示能力。此外,调节影响对抗损失和蒸馏损失之间平衡的超参数也非常重要,需要根据具体任务和模型性能进行调整。因此,探索如何利用师生网络间特征度量空间的差异提高复杂样本的分类和还原效果,提升迁移效率,是值得思考的方向。

3.3　对生成器和判别器蒸馏

对生成器和判别器蒸馏中有两个生成器,即教师生成器和学生生成器。同理,判别器也有教师判别器和学生判别器。其中,教师生成器和教师判别器是 GANs 中原始的生成器和判别器,学生生成器和学生判别器是轻量化后的网络。该类型方法采用在线协同蒸馏策略,不仅聚焦于师生生成器的压缩和优化,同时考虑师生判别器的优化,以保持生成器和判别器之间的平衡,利用双方的中间信息进行协同学习,提升轻量级生成器的性能。

3.3.1　对生成器和判别器蒸馏的实现过程

通过将教师生成器的知识蒸馏到学生生成器,可以指导学生生成器产生更加准确和多样化的数据。一般涉及利用教师网络的输出或特征激活形成一个目标分布或特征表示,学生生成器需要尽可能模仿这个目标。通过将教师判别器的知识蒸馏到学生判别器,可以提高学生判别器区分真伪样本的准确度。这涉及利用教师网络的决策边界或特征表示增强学生判别器的判别能力。利用对生成器和判别器蒸馏的方法不仅可以轻量化 GANs 中原始的生成器和判别器(即教师生成器和教师判别器),还可以改善样本的生成质量和模型的判别能力。对生成器和判别器蒸馏的结构如图 9 所示。

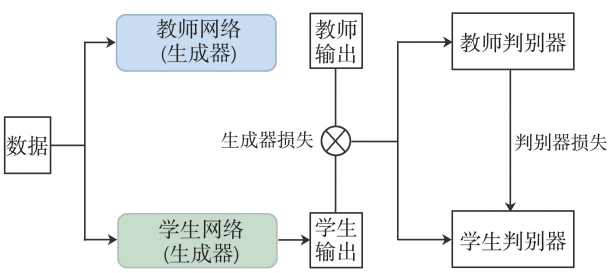

图 9　对生成器和判别器蒸馏的一般示意图

Fig.9　General diagram of distillation against generator and discriminator

在该类型方法中,损失函数由生成器损失和判别器损失构成。生成器损失包括传统对抗损失和蒸馏损失,蒸馏损失可以通过比较生成样本的教师生成器和学生生成器产生的样本输出计算;判别器损失同样包括对抗损失和蒸馏损失,蒸馏损失可以通过比较教师判别器对真实样本和生成样本判决的差异与学生判别器相应差异定义。生成器损失 L_{G} 和判别器损失 L_{D} 的计算式分别为

$$L_{\mathrm{G}}=L_{\mathrm{adv}}+\alpha L_{\mathrm{distill}}^{\mathrm{G}}, \tag{7}$$

$$L_{\mathrm{D}}=L_{\mathrm{adv}}+\mu L_{\mathrm{distill}}^{\mathrm{D}}, \tag{8}$$

式中,L_{adv}为对抗损失,$L_{\mathrm{distill}}^{\mathrm{G}}$为生成器蒸馏损失,

$L_{\text{distill}}^{\text{D}}$为判别器蒸馏损失,$\alpha$ 和 μ 分别为生成器和判别器的平衡系数。

3.3.2 对生成器和判别器蒸馏的实际应用

对生成器和判别器蒸馏的实际应用中,最具代表的工作是文献[72]提出的基于知识蒸馏的生成对抗网络压缩方法,通过继承原教师生成器的低级和高级信息,训练一个参数较少的学生生成器。为了提高学生生成器的能力,该方法引入教师判别器和学生判别器,针对学生网络的生成器和判别器引入不同的蒸馏损失函数,通过学习教师生成器和判别器中蕴含的信息,学生网络可以使用较少的参数取得和教师网络相似的图像转换性能,实现更为便捷的训练和应用。为了优化教师生成器和学生生成器之间的知识传递过程,解决以往学生生成器层仅接收来自教师生成器中相同深度阶段的知识所存在的问题,文献[73]提出一种应用于 cGAN 的累积知识蒸馏(accumulation knowledge distillation, ACKD)方法,充分探索嵌入在中间教师生成器层中的暗知识。文献[74]提出在线对抗特征图蒸馏(online adversarial feature map distillation, AFD)策略,利用对抗训练的方式在线互相学习中间特征图分布,结合原有的样本输出概率学习进一步提升分类精度。在目标检测任务中,文献[75]让教师网络和学生网络分别生成真假样本,通过二者对抗训练提高目标检测性能。

4 基于GANs知识蒸馏的问题与挑战

基于 GANs 的知识蒸馏为模型压缩提供新的研究方向,但仍有以下问题与挑战。

(1)GANs 训练的质量和稳定性。GANs 的训练过程依赖两个网络(生成器和判定器)之间的动态平衡,容易导致训练过程中的多种问题。在知识蒸馏场景下,GANs 的训练不稳定可能导致学生网络学习到不准确或不一致的知识,影响学生网络的性能和泛化能力。此外,不稳定训练可能造成学生网络在学习过程中的收敛问题,增加训练时间和资源消耗。

(2)数据多样性需求高。虽然 GANs 能够生成多样性的数据帮助学生网络学习更加丰富和多样化的数据表示,但是 GANs 在训练过程中存在模式崩溃现象,即生成器开始生成非常相似或完全相同的样本,而不是多样化的样本。这是因为生成器找到一个能够欺骗判别器的“捷径”,无需学习生成真实数据的复杂分布,导致数据多样性不足,影响知识蒸馏过程,使学生网络无法学习多样化的数据。

(3)知识表示的匹配问题。在知识蒸馏过程中,教师网络能否向学生网络传递有效的知识是一个挑战。当教师网络和学生网络架构差异较大时,二者的特征空间会有很大不同。即便生成数据在教师网络中能够有效激活特定的特征表示,也可能不适用于学生网络,因为学生网络的特征提取和处理方式与教师网络有本质区别,学生网络无法从这些特征表示中学到对其有用的信息。

(4)对抗样本的有效性。GANs 用于生成对抗样本以增强学生网络的鲁棒性是一个有前景的研究方向,但生成高质量、有效的对抗样本仍然面临一系列挑战。首先,对抗样本需要在不被人察觉的情况下误导模型。这意味着扰动必须足够细微又足够强大,能够导致模型做出错误预测。设计这样的扰动是极其精细和复杂的,需要精确控制变化程度及扰动方向。其次,过度专注于生成具有普遍性的对抗样本可能会牺牲对特定模型的有效性,反之亦然。因此,找到两者之间的平衡是一个挑战。

(5)计算资源和训练时间。在知识蒸馏的背景下,使用 GANs 生成数据增加训练集的多样性或生成对抗样本,提高学生网络的鲁棒性,但在处理大型模型和数据集时会消耗大量的计算资源和时间成本。为了生成高质量样本,GANs 的生成器需要学习数据的高维分布,使生成器网络需要具有较大的容量和深度,导致模型参数量增加。

(6)可解释性问题。生成对抗网络的知识蒸馏方法通常基于黑盒模型进行,学生网络通过特定的目标函数优化获取教师网络知识,所获知识大多是经验性的结论,无法提供可解释的结果。如何从理论上彻底解释对抗训练过程中知识传递的泛化理论,更好地服务技术发展,是重点研究问题。

上述问题和挑战是基于 GANs 知识蒸馏方法本身的特性,对于特定的应用场景和需求,还需要根据具体情况权衡利弊设计相应的解决方案。

5 未来展望

随着基于结构知识蒸馏方法的发展,基于 GANs 的知识蒸馏对网络轻量化、移动端部署方面有极大作用,未来研究内容可参考如下方面。

(1)优化网络架构和训练技巧。通过引入 GANs 的变体网络(如 Wasserstein 生成对抗网

络[76]、cGAN[77])改进训练策略,缓解训练不稳定性,还可以通过自注意力机制捕捉模型中长距离的依赖关系,提高生成样本的质量[79-80]。此外,训练过程可以采用渐进式、多阶段的训练方法[81-82],如先训练一个 GANs 生成高质量的数据,然后在第二阶段使用这些数据进行知识蒸馏,确保知识蒸馏过程使用的是质量高、多样性好的数据。

(2)避免模式崩溃现象。为了确保生成数据的多样性和覆盖范围,避免生成器出现模式崩溃现象,可以采用多样性鼓励策略鼓励生成器产生彼此不同的样本,还可以采用多个生成器,每个生成器尝试捕获数据分布的不同部分。这些生成器可以并行训练,并与单个判别器或多个判别器竞争。多生成器策略能够覆盖更广泛的数据分布区域,规避因某个生成器出现模式崩溃带来的副作用,有效提高学生网络的学习效果。

(3)借助更灵活的知识蒸馏技术。知识表示的匹配问题是指当教师网络和学生网络架构有较大差异时,如何确保 GANs 生成的数据能够有效表示教师网络知识,并被学生网络有效吸收。可以通过优化知识蒸馏的传递方式解决以上问题。借助中间层知识蒸馏技术,将中间层输出作为额外的蒸馏目标[78]。使用自适应层或转换层(如卷积层、全连接层等)将教师网络的特征映射到一个新的空间,学生网络可以在新空间中更容易地模仿和学习教师网络知识。实施对齐策略,如通过最小化教师和学生网络特征之间的距离,或使用注意力机制加强学生网络对重要特征的关注。

(4)提高对抗样本有效性。为了生成有效的多样化样本,可以采用多目标优化策略,如设计 GANs 的生成器同时优化多个目标,最小化与真实数据的差异和最大化对学生网络的误导性;可以改进对抗过程,如采用动态对抗训练,根据学生网络学习进度的不同,调整对抗样本生成强度和多样性,在训练早期,可以生成较为简单的对抗样本,逐渐增加难度,避免训练初期过度挑战学生网络;可以设计更精细的损失函数,如引入一个专注于增强对抗样本边缘区域的损失函数,使样本在模型的决策边界上具有更高的挑战性,有助于模型学习更加鲁棒的特征表示。

(5)结合其他模型压缩技术。经过基于生成对抗网络知识蒸馏的学生网络存在冗余参数,训练过程不稳定。网络剪枝技术可以移除不重要的参数,实现网络任意程度压缩或变窄,在硬件上能够有效实现加速[83]。参数量化能够替换高精度参数,减少参数存储空间,加快运算速度,降低设备功耗。利用网络剪枝技术和参数量化可以解决网络训练困难的问题,但在缩减模型体量时需要避免出现模型性能衰减。因此,探索如何联合多类型压缩技术实现端到端的快速训练有极高的研究价值。

(6)理论研究的可解释性。现存基于生成对抗网络的知识蒸馏方法主要是经验性结论,缺乏一定的可解释理论,无法从理论上解释哪一种结构性知识更有优势。研究人员可以从现有基础出发,结合传统机器学习和深度学习,从信息论、模型泛化性等角度开展理论研究。

6　结束语

知识蒸馏技术发展至今,与其他交叉领域的应用受到学术界和工业界的广泛关注,并取得许多瞩目的研究成果。知识蒸馏同其他主流技术(如生成对抗网络、强化学习、联邦学习等)的结合通常能够获得意想不到的效果,不仅可以获得轻量级的网络模型,以便应用于资源受限的设备上,还可以通过传递不同形式的知识强化深度学习模型的性能。本研究从知识蒸馏的作用机制、知识传递形式出发,深入探讨 3 种类型的生成对抗网络知识蒸馏技术,并讨论其今后的发展方向。

参考文献:

[1] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. (2015-05-09)[2024-04-20]. https://arxiv.org/abs/1503.02531v1

[2] SUCHOLUTSKY I, SCHONLAU M. Soft-label dataset distillation and text dataset distillation[C]//Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen, China: IEEE, 2021: 1-8.

[3] MALIK S M, HAIDER M U, THARANI M, et al. Teacher-class network: a neural network compression mechanism[EB/OL]. (2021-10-29)[2024-04-30]. https://arxiv.org/abs/2004.03281v3

[4] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 3962-3971.

[5] 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. Washington, D.C., USA: ICLR, 2015: 1-13.

[6] YIM J, JOO D, BAE J, 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 (CVPR). Honolulu, USA: IEEE, 2017: 7130-7138.

[7] BUDNIK M, AVRITHIS Y. Asymmetric metric learning for knowledge transfer [C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021: 8228-8238.

[8] LI X J, WU J L, FANG H Y, et al. Local correlation consistency for knowledge distillation [C]//Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 18-33.

[9] TAO X Y, HONG X P, CHANG X Y, et al. Few-shot class-incremental learning [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 12183-12192.

[10] CHO J H, HARIHARAN B. On the efficacy of knowledge distillation [C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 4793-4801.

[11] ZHOU Z D, ZHUGE C R, GUAN X W, et al. Channel distillation: channel-wise attention for knowledge distillation[EB/OL]. (2020-06-02) [2024-04-25]. https://arxiv.org/abs/2006.01683v1

[12] YUE K Y, DENG J F, ZHOU F. Matching guided distillation[C]//Computer Vision-ECCV 2020. Glasgow, UK: Springer, 2020: 312-328.

[13] CHEN S Y, WANG W Y, PAN S J, et al. Cooperative pruning in cross-domain deep neural network compression[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: ACM, 2019: 2102-2108.

[14] LE D H, VO T N, THOAI N. Paying more attention to snapshots of iterative pruning: improving model compression via ensemble distillation [EB/OL]. (2020-08-14) [2024-04-25]. https://arxiv.org/abs/2006.11487v3

[15] XIE J, LIN S H, ZHANG Y C, et al. Compressing convolutional neural networks with cheap convolutions and online distillation[J]. Displays, 2023, 78: 102428.

[16] XU K R, RUI L, LI Y S, et al. Feature normalized knowledge distillation for image classification [C]// Computer Vision-ECCV 2020. Glasgow, UK: Springer, 2020: 664-680.

[17] CHEN W C, CHANG C C, LEE C R. Knowledge distillation with feature maps for image classification [C]//Computer Vision-ACCV 2018. Perth, Australia: Springer, 2019: 200-215.

[18] HUANG Z Y, ZOU Y, BHAGAVATULA V, et al. Comprehensive attention self-distillation for weakly-supervised object detection[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2020: 16797-16807.

[19] LI M, HALSTEAD M, MCCOOL C. Knowledge distillation for efficient instance semantic segmentation with transformers[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024: 5432-5439.

[20] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM, 2014: 2672-2680.

[21] 张俊三, 程俏俏, 万瑶, 等. MIRGAN: 一种基于GAN的医学影像报告生成模型[J]. 山东大学学报(工学版), 2021, 51(2): 9-18.
ZHANG Junsan, CHENG Qiaoqiao, WAN Yao, et al. MIRGAN: a medical image report generation model based on GAN[J]. Journal of Shandong University (Engineering Science), 2021, 51(2): 9-18.

[22] 张月芳, 邓红霞, 呼春香, 等. 融合残差块注意力机制和生成对抗网络的海马体分割[J]. 山东大学学报(工学版), 2020, 50(6): 76-81.
ZHANG Yuefang, DENG Hongxia, HU Chunxiang, et al. Hippocampal segmentation combining residual attention mechanism and generative adversarial networks [J]. Journal of Shandong University (Engineering Science), 2020, 50(6): 76-81.

[23] GOU J P, YU B S, MAYBANK S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129: 1789-1819.

[24] 黄震华, 杨顺志, 林威, 等. 知识蒸馏研究综述[J]. 计算机学报, 2022, 45(3): 624-653.
HUANG Zhenhua, YANG Shunzhi, LIN Wei, et al. Knowledge distillation: a survey[J]. Chinese Journal of Computers, 2022, 45(3): 624-653.

[25] BUCILUǍ C, CARUANA R, NICULESCU-MIZIL A. Model compression[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA: ACM, 2006: 535-541.

[26] BA L J, CARUANA R. Do deep nets really need to be

deep? [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM, 2014: 2654-2662.

[27] LI J Y, ZHAO R, HUANG J T, et al. Learning small-size DNN with output-distribution-based criteria [C]// Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore: ISCA, 2014: 1910-1914.

[28] TANG Z Y, WANG D, ZHANG Z Y. Recurrent neural network training with dark knowledge transfer [C]// Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China: IEEE, 2016: 5900-5904.

[29] 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. Honolulu, USA: AAAI, 2019: 5628-5635.

[30] YUAN L, TAY F E, LI G L, et al. Revisiting knowledge distillation via label smoothing regularization [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 3902-3910.

[31] ZHAO B R, CUI Q, SONG R J, et al. Decoupled knowledge distillation [C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE, 2022: 11943-11952.

[32] XIE Q Z, LUONG M T, HOVY E, et al. Self-training with noisy student improves ImageNet classification [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 10684-10695.

[33] GUPTA S, HOFFMAN J, MALIK J. Cross modal distillation for supervision transfer [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 2827-2836.

[34] KOOHPAYEGANI A S, TEJANKAR A, PIRSIAVASH H. CompRess: self-supervised learning by compressing representations [EB/OL]. (2020-10-28) [2024-4-25]. https://arxiv.org/abs/2010.14713v1

[35] YUN S, PARK J, LEE K, et al. Regularizing class-wise predictions via self-knowledge distillation [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 13873-13882.

[36] WU G L, GONG S G. Peer collaborative learning for online knowledge distillation [EB/OL]. (2021-03-03) [2024-4-25]. https://arxiv.org/abs/2006.04147v2

[37] 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 (ICCV). Seoul: IEEE, 2019: 5007-5016.

[38] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer [EB/OL]. (2017-02-12) [2024-04-25]. https://arxiv.org/abs/1612.03928v3

[39] TUNG F, MORI G. Similarity-preserving knowledge distillation [C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 1365-1374.

[40] PASSALIS N, TEFAS A. Learning deep representations with probabilistic knowledge transfer [C]//Computer Vision-ECCV 2018. Munich, Germany: Springer, 2018: 283-299.

[41] GUAN Y S, ZHAO P Y, WANG B X, et al. Differentiable feature aggregation search for knowledge distillation [C]//Computer Vision-ECCV 2020. Glasgow, UK: Springer, 2020: 469-484.

[42] LEE S H, KIM D H, SONG B C. Self-supervised knowledge distillation using singular value decomposition [C]//Computer Vision-ECCV 2018. Munich, Germany: Springer, 2018: 339-354.

[43] HEO B, LEE M, YUN S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons [C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Palo Alto, USA: ACM, 2019: 3779-3787.

[44] KIM J, PARK S, KWAK N, et al. Paraphrasing complex network: network compression via factor transfer [C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: ACM, 2018: 2765-2774.

[45] AHN S, HU S X, DAMIANOU A, et al. Variational information distillation for knowledge transfer [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 9163-9171.

[46] YIM J, JOO D, BAE J, 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 (CVPR). Honolulu, USA: IEEE, 2017:7130-7138.

[47] SRINIVAS S, FLEURET F. Knowledge transfer with Jacobian matching[EB/OL]. (2018-03-01)[2024-04-25]. https://arxiv.org/abs/1803.00443v1

[48] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 3962-3971.

[49] LASSANCE C, BONTONOU M, HACENE G B, et al. Deep geometric knowledge distillation with graphs[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 8484-8488.

[50] LIU Y F, CAO J J, LI B, et al. Knowledge distillation via instance relationship graph[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2019: 7089-7097.

[51] XU X X, ZOU Q, LIN X, et al. Integral knowledge distillation for multi-person pose estimation[J]. IEEE Signal Processing Letters, 2020, 27: 436-440.

[52] 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 (CVPR). Seattle, USA: IEEE, 2020: 12483-12492.

[53] CHEN X, ZHANG Y F, XU H T, et al. Adversarial distillation for efficient recommendation with external knowledge[J]. ACM Transactions on Information Systems, 2018, 37(1): 1-28.

[54] XU Z, HSU Y C, HUANG J. Training student networks for acceleration with conditional adversarial networks[C]//Proceedings of the 2018 British Machine Vision Conference (BMVC). Newcastle, UK: BMVA, 2018:61.

[55] ZHANG T C, LIU Y X. MTUW-GAN: a multi-teacher knowledge distillation generative adversarial network for underwater image enhancement[J]. Applied Sciences, 2024, 14(2): 529.

[56] WANG Y H, XU C, XU C, et al. Adversarial learning of portable student networks[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: ACM, 2018: 4260-4267.

[57] BELAGIANNIS V, FARSHAD A, GALASSO F. Adversarial network compression[C]// Computer Vision-ECCV 2018. Munich, Germany: Springer, 2019: 431-449.

[58] AGUINALDO A, CHIANG P Y, GAIN A, et al. Compressing GANs using knowledge distillation[EB/OL]. (2019-02-01)[2024-4-30]. https://arxiv.org/abs/1902.00159v1

[59] WANG X, ZHANG R, SUN Y, et al. KDGAN: knowledge distillation with generative adversarial networks[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: ACM, 2018: 783-794.

[60] REN Y X, WU J, XIAO X F, et al. Online multi-granularity distillation for GAN compression[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE, 2021: 6773-6783.

[61] HU T, LIN M B, YOU L Z, et al. Discriminator-cooperated feature map distillation for GAN compression[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 20351-20360.

[62] VO D M, SUGIMOTO A, NAKAYAMA H. PPCD-GAN: progressive pruning and class-aware distillation for large-scale conditional GANs compression[C]// Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2022: 1422-1430.

[63] CHEN H T, WANG Y H, XU C, et al. Data-free learning of student networks[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 3513-3521.

[64] YE J W, JI Y X, WANG X C, et al. Data-free knowledge amalgamation via group-stack dual-GAN[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2020: 12513-12522.

[65] 张晶, 鞠佳良, 任永功. 基于双生成器网络的 Data-Free 知识蒸馏[J]. 计算机研究与发展, 2023, 60(7): 1615-1627.
ZHANG Jing, JU Jialiang, REN Yonggong. Double-generators network for Data-Free knowledge distillation[J]. Journal of Computer Research and Development, 2023, 60(7): 1615-1627.

[66] MICAELLI P, STORKEY A. Zero-shot knowledge transfer via adversarial belief matching[EB/OL]. (2019-11-25)[2024-04-30]. https://arxiv.org/abs/

1905.09768v4

[67] FANG G F, SONG J, SHEN C C, et al. Data-free adversarial distillation[EB/OL]. (2020-03-02)[2024-04-30]. https://arxiv.org/abs/1912.11006v3

[68] YU X Y, YAN L, YANG Y, et al. Conditional generative data-free knowledge distillation[J]. Image and Vision Computing, 2023, 131: 104627.

[69] DO K, LE T H, NGUYEN D, et al. Momentum adversarial distillation: handling large distribution shifts in data-free knowledge distillation[EB/OL]. (2022-09-21)[2024-04-30]. https://arxiv.org/abs/2209.10359v1

[70] CHOI Y, CHOI J, EL-KHAMY M, et al. Data-free network quantization with adversarial knowledge distillation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2020: 710-711.

[71] CUI K W, YU Y C, ZHAN F N, et al. KD-DLGAN: data limited image generation via knowledge distillation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023: 3872-3882.

[72] CHEN H T, WANG Y H, SHU H, et al. Distilling portable generative adversarial networks for image translation[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI, 2020: 3585-3592.

[73] GAO T W, LONG R J. Accumulation knowledge distillation for conditional GAN compression[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE, 2023: 1294-1303.

[74] CHUNG I, PARK S, KIM J, et al. Feature-map-level online adversarial knowledge distillation[EB/OL]. (2020-06-05)[2024-04-30]. https://arxiv.org/abs/2002.01775v3

[75] WANG W W, HONG W, WANG F, et al. GAN-knowledge distillation for one-stage object detection[J]. IEEE Access, 2020, 8: 60719-60727.

[76] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]//Proceedings of the 2017 International Conference on Machine Learning (ICML). Sydney, Australia: JMLR, 2017: 214-223.

[77] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2024-04-30]. https://arxiv.org/abs/1411.1784v1

[78] CHEN P G, LIU S, ZHAO H S, et al. Distilling knowledge via knowledge review[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE, 2021: 5006-5015.

[79] HUANG Z Z, LIANG M F, QIN J H, et al. Understanding self-attention mechanism via dynamical system perspective[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 1412-1422.

[80] 兰治, 严彩萍, 李红, 等. 混合双注意力机制生成对抗网络的图像修复模型[J]. 中国图象图形学报, 2023, 28(11): 3440-3452.
LAN Zhi, YAN Caiping, LI Hong, et al. HDA-GAN: hybrid dual attention generative adversarial network for image inpainting[J]. Journal of Image and Graphics, 2023, 28(11): 3440-3452.

[81] 黄仲浩, 杨兴耀, 于炯, 等. 基于多阶段多生成对抗网络的互学习知识蒸馏方法[J]. 计算机科学, 2022, 49(10): 169-175.
HUANG Zhonghao, YANG Xingyao, YU Jiong, et al. Mutual learning knowledge distillation based on multi-stage multi-generative adversarial network[J]. Computer Science, 2022, 49(10): 169-175.

[82] 钱亚冠, 马骏, 何念念, 等. 面向边缘智能的两阶段对抗知识迁移方法[J]. 软件学报, 2022, 33(12): 4504-4516.
QIAN Yaguan, MA Jun, HE Niannian, et al. Two-stage adversarial knowledge transfer for edge intelligence[J]. Journal of Software, 2022, 33(12): 4504-4516.

[83] SHI Y, TANG A D, NIU L F, et al. Sparse optimization guided pruning for neural networks[J]. Neurocomputing, 2024, 574: 127280.

(编辑:孙亚彤)