深度学习中知识蒸馏研究综述

邵仁荣　　刘宇昂　　张伟　　王骏

(华东师范大学 计算机科学与技术学院, 上海 200062)

摘　要　在人工智能迅速发展的今天，深度神经网络广泛应用于各个研究领域并取得了巨大的成功，但也同样面临着诸多挑战。首先，为了解决复杂的问题和提高模型的训练效果，模型的网络结构逐渐被设计的深而复杂，难以适应移动计算发展对低资源、低功耗的需求。知识蒸馏最初作为一种从大型教师模型向浅层学生模型迁移知识、提升性能的学习范式被用于模型压缩。然而随着知识蒸馏的发展，其教师-学生的架构作为一种特殊的迁移学习方式，演化出了丰富多样的变体和架构，并被逐渐扩展到各种深度学习任务和场景中，包括计算机视觉、自然语言处理、推荐系统等等。另外，通过神经网络模型之间迁移知识的学习方式，可以联结跨模态或跨域的学习任务，避免知识遗忘；还能实现模型和数据的分离，达到保护隐私数据的目的。知识蒸馏在人工智能各个领域发挥着越来越重要的作用，是解决很多实际问题的一种通用手段。本文将近些年来知识蒸馏的主要研究成果进行梳理并加以总结，分析该领域所面临的挑战，详细阐述知识蒸馏的学习框架，从多种分类角度对知识蒸馏的相关工作进行对比和分析，介绍了主要的应用场景，在最后对未来的发展趋势提出了见解。

关键词　深度神经网络；知识蒸馏；模型压缩；迁移学习；人工智能

中图法分类号 TP391

A Survey of Knowledge Distillation in Deep Learning

SHAO Ren-Rong　　LIU Yu-Ang　　ZHANG Wei　　WANG Jun

(School of Computer Science and Technology, East China Normal University, 200062, Shanghai)

Abstract　With the rapid development of artificial intelligence, deep neural networks are widely used in various research fields and have achieved great success, but they also face a lot of challenges. First of all, to solve more complex problems and improve the training effect of the model, the network structure of the model is gradually designed to be deep and complex, and it is difficult to adapt to the development of mobile computing for low resources and low power consumption. Knowledge distillation was originally used for model compression as a learning paradigm that transfers knowledge from a large teacher model to a compact student model and improves performance. However, with the development of knowledge distillation, its teacher-student architecture, as a special transfer learning method, has evolved a rich variety of variants and architectures, and has been gradually extended to various deep learning tasks and scenarios, including computers vision, natural language processing, recommendation systems, etc. In addition, through the learning method of transferring knowledge between neural network models, cross-modal or cross-domain learning tasks can be connected to avoid knowledge forgetting; it can also achieve the separation of models and data to achieve the purpose of protecting private data. Knowledge distillation is playing an increasingly important role in various fields of artificial intelligence, and it is a universal means to solve many practical problems. This paper sorts out the main references of knowledge distillation, elaborates the learning framework of knowledge distillation, compares and analyzes the related work

本课题得到国家自然科学基金(No. 62072182)资助. **邵仁荣**，博士研究生，计算机学会(CCF)会员（G2938G），主要研究领域为计算机视觉、模型压缩. E-mail: roryshao@foxmail.com. **刘宇昂（共同一作）**，博士研究生，主要研究领域为知识蒸馏、模型压缩、计算机视觉. E-mail: frankliu624@gmail.com. **张 伟（通信作者）**，博士，副研究员，计算机学会(CCF)会员（62154M），主要研究领域为推荐系统、数据挖掘、知识蒸馏. E-mail: zhangwei.thu2011@gmail.com. **王 骏（通信作者）**，博士，教授，主要研究领域为计算机视觉、机器学习. E-mail: wongjun@gmail.com.

of knowledge distillation from a variety of classification perspectives, introduces the main application scenarios, and finally discusses the future development trends and provides insights.

Key words Deep Neural Network; Knowledge Distillation; Model Compression; Transfer Learning; Artificial Intelligence

1 引言

随着深度神经网络的崛起和演化，深度学习在计算机视觉[1-3]、自然语言处理[4-6]、推荐系统[7-9]等各个人工智能的相关领域中已经取得了重大突破。但是，深度学习在实际应用过程中的也存在着一些巨大的挑战。首先，为了应对错综复杂的学习任务，深度学习的网络模型往往会被设计的深而复杂：比如早期的 LeNet 模型只有 5 层，发展到目前的通用的 ResNet 系列模型已经有 152 层；伴随这模型的复杂化，模型的参数也在逐渐加重。早期的模型参数量通常只有几万，而目前的模型参数动辄几百万。这些模型的训练和部署都需要消耗大量的计算资源，且模型很难直接应用在目前较为流行的嵌入式设备和移动设备中。其次，深度学习应用最成功的领域是监督学习，其中很多任务上的表现几乎已经超越了人类的表现。但是，监督学习需要依赖大量的人工标签；而要实现大规模的标签任务是非常困难的事情，一方面是数据集的获取，在现实场景中的一些数据集往往很难直接获取。比如，在医疗行业需要保护患者的隐私数据，因而数据集通常是不对外开放的。另一方面，大量的用户数据主要集中在各个行业的头部公司的手中，一些中小型公司无法积累足够多的真实用户数据，因此模型的效果往往是不理想的；此外，标注过程中本身就需要耗费很大的人力、物力、财力，这将极大限制人工智能在各个行业中的发展和应用。最后，从产业发展的角度来看，工业化将逐渐过渡到智能化，边缘计算逐渐兴起预示着 AI 将逐渐与小型化智能化的设备深度融合，这也要求模型更加的便捷、高效、轻量以适应这些设备的部署。

针对深度学习目前在行业中现状中的不足，Hinton 于 2015 首次提出了知识蒸馏(Knowledge Distillation, KD)[10]，利用复杂的深层网络模型向浅层的小型网络模型迁移知识。这种学习模型的优势在于它能够重用现有的模型资源，并将其中蕴含的信息用于指导新的训练阶段；在跨领域应用中还改变了以往任务或场景变化都需要重新制作数据集和训练模型的困境，极大地节省了深度神经网络训练和应用的成本。通过知识蒸馏不仅能够实现跨领域和跨模态数据之间的联合学习还能将模型和知识表示进行分离，从而在训练过程中将教师模型作为“黑盒”处理，可以避免直接暴露敏感数据，达到隐私保护效果。

知识蒸馏作为一种新兴的、通用的模型压缩和迁移学习架构，在最近几年展现出蓬勃的活力，其发展历程也大致经历了初创期，发展期和繁荣期。在初创期，知识蒸馏从输出层逐渐过渡到中间层，这时期知识的形式相对简单，较为代表性的中间层特征蒸馏的方法为 Hints[11]。到了发展期，知识的形式逐渐丰富、多元，不在局限于单一的节点，这一时期较为代表性的蒸馏方法有 AT[12]，FT[13]。在 2019 年前后，知识蒸馏逐渐吸引了深度学习各个领域研究人员的目光，使其应用得到了广泛拓展，比如在模型应用上逐渐结合了跨模态[14-17]、跨领域[18-21]、持续学习[22-24]、隐私保护[25,26]等；在和其它领域交叉过程中又逐渐结合了对抗学习[27,28]、强化学习[29,30]、元学习[31-33]、自动机器学习[34-36]、自监督学习[37,38]等。如下图 1 为知识蒸馏的发展历程和各个时期较为代表性的工作。

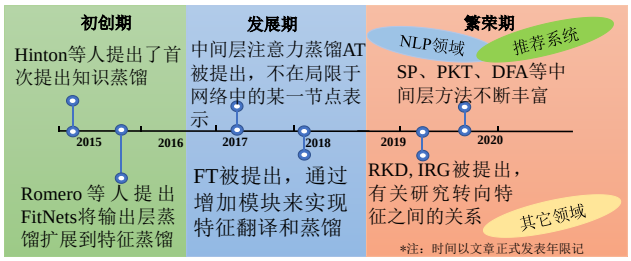


图 1 知识蒸馏发展历程

知识蒸馏虽然有了较为广阔的发展，但是在其发展过程和实际应用中也同样面临的这一些挑战；知识蒸馏的挑战主要可以分为实际应用中面临的挑战和模型本身理论上的挑战。应用中的挑战主要有模型问题，成本问题；而理论上面存在的主要挑战也是目前深度学习普遍存在的一些挑战，包括模型的不可解释性等：

模型问题：在实际工业应用中针对不同的任务教师模型多样，而如果教师和学生模型不匹配，可能会使学生模型无法模仿深层大容量的教师模型，即大模型往往不能成为更好的老师。因此，应用中需要考虑深层模型和浅层模型之间的容量差距，选

择相匹配的教师-学生模型。

成本问题：模型训练过程对超参数较为敏感以及对损失函数依赖较大，而相关原因很难用原理去解释，需要大量的实验，因而模型的试错成本相对较高。

可解释性不足：关于知识蒸馏的原理解释主要是从输出层标签平滑正则化、数据增强等角度出发，而关于其他层的方法原理解释相对不足；目前，虽然关于泛化边界的研究也在兴起，但是并不能全面解释知识的泛化问题，还需要有更进一步的探究，才能保证理论的完备性。

目前，知识蒸馏已经成为一个热门的研究课题，关于知识蒸馏的论文和研究成果非常丰富。各种新方法、新任务、新场景下的研究纷繁复杂，使得初学者难以窥其全貌。当前已有两篇关于知识蒸馏的综述[39,40]，均发表于 2021 年。相较于前者[39]，本文在分类上作了进一步细化，如在知识形式上，本文关注到了参数知识及蒸馏中常见的中间层的同构和异构问题；虽然该文献中也提及了基于图的算法，但是本文以为基于图形式构建的知识表示是一种新兴的、独立的、特殊的知识形式，单独归为一类更为合理。相较于后者[40]本文在结构分类上更加宏观，以知识形式、学习方式和学习目的为主要内容将知识蒸馏的基础解析清楚，而后在此基础之上对其交叉领域和主要应用进行展开。本文的主要贡献可总结如下：

1) 结构较为完善，分类更加细化。对于知识的分类，本文是依据知识蒸馏的发展脉络对其进行归类并细化，增加了中间层知识、参数知识、图表示知识，完整的涵盖了目前知识的全部形式。在文章的结构上，既保证了分类的综合性，又避免了分类过多分类造成的杂糅，更为宏观。

2) 对比详细，便于掌握。本文以表格的方式对同的方法之间的优缺点、适用场景等进行详细的总结对比，以及对比了不同知识形式蒸馏的形式化方法，使得读者能够快速准确地区分其中的不同点。

3) 内容完整，覆盖全面。本文遵循了主题式分类原则不仅分析了单篇文献，还分析相关领域中知识蒸馏的重要研究。除此之外，本文以独立章节对知识蒸馏的学习目的，原理和解释，发展趋势等方面做了较为全面的阐释。

本文接下来将从知识蒸馏的整体框架从发，并对其各个分类进行详细的阐述，使得读者能够从宏观上对知识蒸馏有更全面的了解，以便更好地开展相关领域的学习与研究。本文将按照以下结构组织：

第 2 章首先介绍了知识蒸馏的理论的基础及分类。第 3-6 章分别按照知识传递形式、学习方式、学习目的、交叉领域的顺序，从 4 个不同角度对知识蒸馏的相关工作进行了分类和对比，并分析了不同研究方向面临的机遇和挑战; 第 7 章列举了知识蒸馏在计算机视觉、自然语言处理、推荐系统等领域的一些应用性成果。第 8 章对知识蒸馏的原理和可解释性方面的工作进行了梳理。最后，对知识蒸馏在深度学习场景下的未来发展趋势提出了一些见解，并进行全文总结。

2 理论基础及分类

知识蒸馏本质上属于迁移学习的范畴，其主要思路是将已训练完善的模型作为教师模型，通过控制“温度”从模型的输出结果中“蒸馏”出“知识”用于学生模型的训练，并希望轻量级的学生模型能够学到教师模型的“知识”，达到和教师模型相同的表现。这里的“知识”狭义上的解释是教师模型的输出中包含了某种相似性，这种相似性能够被用迁移并辅助其它模型的训练，文献[10]称之为“暗知识”；广义上的解释是教师模型能够被利用的一切知识形式，如特征、参数、模块等等。而“蒸馏”是指通过某些方法（如控制参数），能够放大这种知识的相似性，并使其显现的过程；由于这一操作类似于化学实验中“蒸馏”的操作，因而被形象地称为“知识蒸馏”。

从知识蒸馏模型演化来看，“知识”最先是在教师模型输出层实现的。通常，模型输出的 logits 代表着模型对该类别的“判断”，这些判断需要经过 softmax 层后得到类别的预测概率，而后通过样本真实标签(Ground Truth) 直接计算模型的损失。但是，这样的类别概率相对来说都是硬目标(Hard Target)。输出的概率中包含的类别之间相似性信息被忽略掉了，这在很大程度上影响了模型的泛化能力。Hinton[10]认为输出的信息中包含了类别之间的相似性“暗知识”(Dark Knowledge)，这些相似性信息也同样具有价值[10]。因此，通过引入温度 T 方法来软化 softmax 输出分类信息。如式 1 所示：

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

其中，z_i、z_j为 logits 作为 softmax 的输入;q_i

对应着每个类别的输出概率，T 代表温度系数，可以控制输出概率的软化 (Soft) 程度。当 $T=1$ 时，公式退化为 softmax 函数，当 T 取值较大时就会得到一个较为软化的概率分布。比如在手写字符识别中数字“2”和数字“7”在外形上是相似的，因此如果使用硬目标来表示数字会得到如图 2(a) 中“非 0 即 1”的表示。但是，引入 T 后软化的分类输出中会得到每一个类别的概率表示如图 2(b)。实验证明软化后的类别分布更易于提升学生模型的学习效果。

在蒸馏框架中教师模型一般是预训练模型，其模型结构和设计通常较为复杂，具有很好的学习、

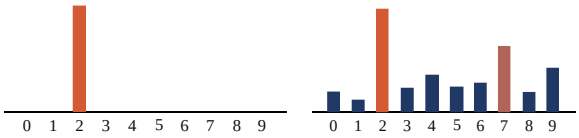


图 2 (a) 左侧表示网络直接输出的硬标签，(b) 右侧经过软化后的标签，能够表示出分类信息中的“暗知识”

表示和泛化能力。输出层蒸馏主要是将学习到的输出层知识按公式 (1) 的方式软化，并作为学生的学习目标；学生模型在训练时，将学习到的分类预测软化来拟合教师模型，同时也将未经软化的分类预测与样本标签进行拟合。教师-学生网络经过 softmax 后得到的概率分布代表着各自对分类任务预测的信息分布。因此，通常采用优化两者之间的交叉熵(Cross Entropy) 损失来衡量知识迁移的目标来引导学生学习并拟合教师模型的概率分布。

$$p_i^T=\frac{\exp(z_i/T)}{\sum_j^N \exp(z_j/T)},\quad q_i^T=\frac{\exp(z_i/T)}{\sum_k^N \exp(z_k/T)} \tag{2}$$

这里 p_i^T，q_i^T分别表示教师和学生模型经过“温度”T 蒸馏后的概率分布。

$$L_{soft}=-\sum_j^N p_j^T \log(q_j^T),\quad L_{hard}=-\sum_j^N c_j \log(q_j^1) \tag{3}$$

其中，c_j 为真实标签，q_j^1 表示温度值为 $T=1$ 的情况。整个模型的损失通常由学生模型预测值和真实值之间的损失$L_{h\,ard}$以及教师模型和学生模型蒸馏后的交叉熵损失L_{soft}构成，如下公式(4)：

$$L=\alpha L_{soft}+(1-a)L_{hard} \tag{4}$$

其中，α是平衡因子。

实际应用中，学生模型的训练效果以及泛化能力受限于模型和设备。因此，设计一个性能优秀的教师模型或者通过预训练教师模型预测样本，可以得到泛化能力较强的输出结果，再将其软化后的 logits 作为知识直接转移给小容量模型，让小容量学生模型来拟合教师模型的分布，以提升学生模型的泛化能力。因此，与知识蒸馏相结合的模型框架基本上遵循了教师模型和学生模型相结合的架构设计。“知识”也由最初的输出层蒸馏逐渐转向中间层蒸馏，其表现形式也朝着结构化和图表示的方向发展和演化。

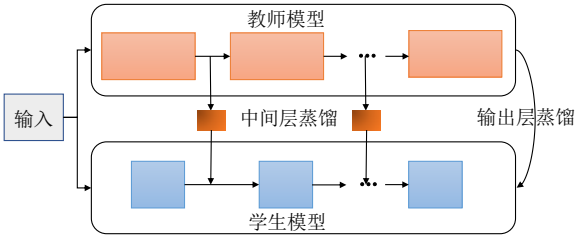


图 3 知识蒸馏教师学生模型结构流程图

如图 3 是知识蒸馏模型的整体结构。其由一个多层的教师模型和学生模型组成，教师模型主要负责向学生模型传递知识，这里的“知识”包括了标签知识、中间层知识、参数知识、结构化知识、图表示知识。在知识的迁移过程中，通过在线或离线等不同的学习方式将“知识”从教师网络转移到了学生网络。为了便于读者快速学习和对比其中的差异，作者将不同知识传递形式下的蒸馏方法的形式化表示及其相关解释整理为下表 1 所示结果。此外，本文对知识蒸馏相关研究进行了总结，主要从知识传递形式、学习的方式、学习的目的、交叉领域、主要应用等方面对其进行分类，其分类框架如图 4 所示，具体内容将在后续的文章中展开。

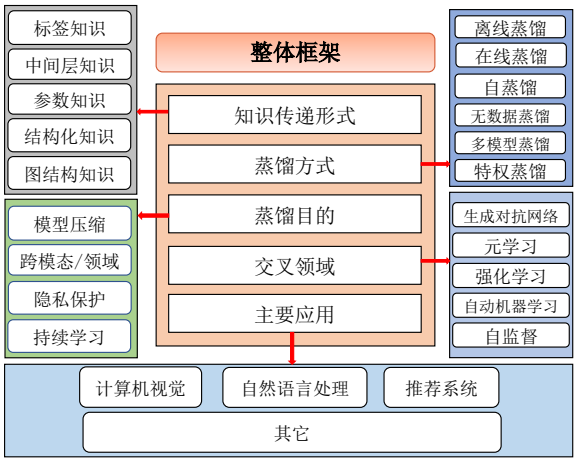


图 4 知识蒸馏整体分类框架

表 1 不同知识传递形式下的蒸馏方法形式化表示对比表

知识形式	损失函数	解释
标签知识	$\mathrm{L_{KD}}=\mathrm{H}\left(\mathbf{y}_{\text{true}},\mathrm{P}_S\right)+\lambda\mathrm{H}\left(\mathrm{P}_T^{\tau},\mathrm{P}_S^{\tau}\right)$	H 为交叉熵，λ 为平衡因子。$\mathbf{y}_{\text{true}}$ 为样本的标签知识，P_S 为学生模型经过 softmax 输出的概率分布，P_T^{τ} 和 P_S^{τ} 为教师和学生模型，在‘温度’为 τ 时 softmax 的概率分布。
中间层知识	$\mathrm{L_{KD}}=\left\|f_T\left(\mathbf{x};\mathbf{W}_T\right)-f_S\left(\mathbf{x};\mathbf{W}_S\right)\right\|_F$	**同构蒸馏**：f_T 和 f_S 分别为教师和学生模型，$\mathbf{x}$ 为中间层输入的特征；$\mathbf{W}_T$ 和 $\mathbf{W}_S$ 为教师和学生模型的中间层参数，F 为范数值。
	$\mathrm{L_{KD}}=\left\|f_S\left(\mathbf{x};\mathbf{W}_S\right)-\gamma\left(f_T\left(\mathbf{x};\mathbf{W}_T\right),\mathbf{W}_R\right)\right\|_F$	**异构蒸馏**：参数符号同上表同构蒸馏，γ 为额外的特征适配器，$\mathbf{W}_R$ 为适配器训练参数。
参数知识	$\mathrm{L}=\underset{\{\mathbf{x}_i\}_{i=1}^{n_l+n_u}}{\mathrm{E}}D\left(f\left(\mathbf{x}_i,\xi';\theta'\right),f\left(\mathbf{x}_i,\xi'';\theta''\right)\right)$ 其中，$\theta_t^{''}=\alpha\theta_{t-1}^{'}+(1-\alpha)\theta_t^{'}$	**教师平均**：ξ' 和 ξ'' 是随机分布，D 是衡量两个模型差异度量函数，通常是均方误差，或者 KL 散度，θ'' 是当前教师模型的权重。t 是当前训练次数，α 是平衡因子。
	$\mathrm{L}\left(\mathbf{Q}_j\right)=\sum_{j=1}^{M}\sum_{i=1}^{N}\left\|\mathbf{Q}_j*\mathrm{Y}_{ij}^{S}-\mathrm{Y}_{ij}^{T}\right\|_F$ 其中，$\mathrm{Y}^S,\mathrm{Y}^T\in\mathrm{R}^{n_o\times d}$	**模块注入**：$\mathbf{Q}_j$ 是对应于卷积层中第 j 层的张量，Y_{ij}^{S} 和 Y_{ij}^{T} 对应学生模型和教师模型模块的输出。
结构化知识	$\mathrm{L_{SKD}}=\sum_{(x_i,..,x_n)\in\mathrm{X}^n}D\left(\psi\left(t_1,..,t_n\right),\psi\left(s_1,..,s_n\right)\right)$	其中，$\left(x_1,\cdots,x_n\right)$ 表示 n 元组特征，ψ 表示特征之间的关系函数，D 表示距离度量。
图表示知识	$\mathrm{L_{GKD}}=\sum_{\ell\in\Lambda}D\left(\mathrm{G}_\ell^S(\mathbf{x}),\mathrm{G}_\ell^T(\mathbf{x})\right)$ 其中，$\mathrm{G}_\ell(\mathbf{x})=\left\langle\mathbf{x}_\ell,\mathbf{W}_\ell\right\rangle$	G_ℓ^S 和 G_ℓ^T 为学生和教师在 ℓ 层上的图结构构造函数，$\mathbf{x}_\ell$ 为输入节点，$\mathbf{W}_\ell$ 表示边的权重矩阵。

3 知识传递形式

知识蒸馏方法的核心在于“知识”的设计、提取和迁移方式的选择，通常不同类型的知识来源于网络模型不同组件或位置的输出。根据知识在教师-学生模型之间传递的形式可以将其归类为标签知识、中间层知识、参数知识、结构化知识和图表示知识。标签知识一般指在模型最后输出的 logits 概率分布中的软化目标信息；中间层知识一般是在网络中间层输出的特征图中表达的高层次信息；参数知识是训练好的教师模型中存储的参数信息；结构化知识通常是考虑多个样本之间或单个样本上下文的相互关系；图表示知识一般是将特征向量映射至图结构来表示其中的关系，以满足非结构化数据表示的学习需求。本章节主要对蒸馏知识的 5 类传递形式加以介绍，理清主流的知识蒸馏基础方法，后续所介绍的各类蒸馏方法或具体应用都是以此为基础。相关形式，优缺点和实验对比，见表 2、3、4。

3.1 标签知识

标签知识是神经网络对样本数据最终的预测输出中包含的潜在信息，这也是目前蒸馏过程中最简单、应用最多的方式。Hinton 等人[10]最早提出的知识蒸馏方法就属于此类。由于经过“蒸馏温度”调节后的软标签中具有很多不确定信息，通常的研究[10,41,42] 认为这其中反映了样本间的相似度或干扰性、样本预测的难度，因此标签知识又被称为“暗知识”。

虽然标签知识通常提供的信息十分有限且有相对的不确定性，但它仍然是基础蒸馏方法研究的重点和热点之一，因为其与传统的伪标签学习[38-41]或者自训练(Self-training)[43,44]方法有着密切的联系，这实际上为半监督学习开辟了新的道路。为了有效地解决基于聚类的算法中的伪标签噪声的问题，Ge 等人[45]利用“同步平均教学”的蒸馏框架进行伪标签优化，核心思想是利用更为鲁棒的“软”标签对伪标签进行在线优化。类似地，MLP[46]提出了基于元学习(Meta-learning) 自适应生成目标分布的方法，用于教师和学生模型的伪标签学习过程。

利用一个筛选网络从目标检测模型预测的伪标签中区分出正例和负例，将正例用于下一阶段的半监督自训练过程，可以有效提升标签数据的利用率[43]。Xie 等人[44]利用有监督训练学生模型自身，在自蒸馏训练中额外地引入无标签噪声数据产生伪标签，将 ImageNet 的 Top-1 识别结果提高了约 1%。对于标签知识蒸馏方法本身，已经有非常多的变体和应用，主要是从改进蒸馏过程、挖掘标签信息、去除干扰等方面，提升学生模型的性能。Gao 等人[47]实现了一种简单的逐阶段的标签蒸馏训练过程，在梯度下降训练过程中，每次只更新学生网络的一个模块，从前至后直到全部更新完成。根据 Mirzadeh 等人[48]的研究发现，并不是教师模型性能越高对于学生模型的学习越有利，当教师-学生模型之间的差距过大时，会导致学生难以从教师模型获得提升。为此，他们提出使用辅助教师策略来逐渐缩小教师和学生之间的学习差距，取得更好的蒸馏效果。同样是为了缩小教师-学生之间的学习差距，Yang 等人[49]则提出利用教师模型在每个训练周期更新的中间模型产生的标签知识指导学生模型。为了充分挖掘标签信息、去除干扰，Muller 等人[50]采用了子类别蒸馏方法，将原标签分组合并参与软标签蒸馏学习；文献[51]则研究了蒸馏损失函数对L_2范数和归一化的软标签的作用，提出使用球面空间度量蒸馏的方法去除范数的影响；Zhang 等人[52]关注了样本权重的影响，通过预测不确定性自适应分配样本权重，改善蒸馏过程；Wu 等人[53]提出了同伴协同蒸馏，通过训练多个分支网络并将其他训练较强教师的 logits 知识转移给同伴，有利于模型的稳定和提高蒸馏的质量。

使用标签知识蒸馏学习最大的优势在于，它无需关注神经网络模型的内在结构或特征表达，直接利用模型对样本的预测输出。这无论是在一般的有监督学习任务，还是跨领域[54]、跨模态[55]的学习任务中，甚至是多模型学习[56-61]、自蒸馏[62-67]、自监督学习[37,38,68-75]等特殊场景中，都是非常简便有效的方式。并且，标签知识蒸馏可以与其他的蒸馏学习方式组合使用，不需要任何额外的设计。

关于“暗知识”的蒸馏作用原理始终是学者们密切关心的问题，但是由于神经网络本身的可解释性研究有限，导致对软标签蒸馏的原理分析十分困难，本文将会在第 8 节中探讨关于蒸馏原理的解释。但是大量的实验性研究表明，所谓的“暗知识”其实是软化的标签对网络的学习产生的正则化效用[41,62,76]。因而，通过标签平滑正则化(Label Smoothing Regularization, LSR)[62]可以在一定程度上模拟出知识蒸馏的效果。最近，笔者[65,77]提出了轻量化的知识蒸馏框架，它首先在一个合成的简单数据集上训练一个轻量教师网络，其类别数与目标数据集的类别数相等，然后教师产生软目标，将 KD 损失和对抗性损失的结合作为增强 KD 损失，可以引导学生学习，起到了利用轻量化辅助教师模型的正则化输出作为一种先验分布的效果。

标签知识蒸馏方法虽然简单易用并且提升效果显著，但其局限性在于模型输出的标签预测表达的信息有限且只适合分类任务，通常需要和其他方式结合起来使用。目前大部分标签知识蒸馏的基础研究和改进还是集中于传统的分类任务，关于如语义分割[78-82]、目标检测[1-3,83]、行人重识别(ReID)[84]等高级视觉任务的标签知识挖掘上研究有限。它们一般是单样本多标签的预测任务，需要考虑标签之间的上下文关系和样本不平衡性，并且使用手工设定的“温度”控制软标签的方式不够灵活，难以适应多目标对象的密集预测任务。对于标签蒸馏的原理虽然有很多实验性证明，但是缺乏严格的数学推导，也无法回归到玻尔兹曼机解释，这一方面的研究可能需要数学和神经网络机制方面的突破来推动。

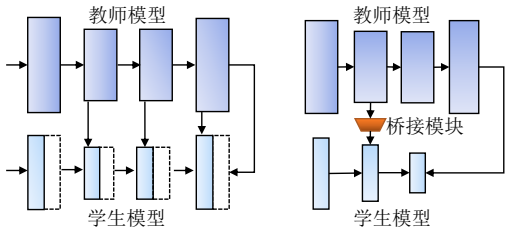
3.2 中间层知识

由于输出层的标签知识提供的信息有限，因此有研究人员希望可以在中间层获取更具表征能力的特征知识转移给学生模型[11]。中间层知识所表达的是深度神经网络的中间层部件所提取出的高维特征。借助外部的知识指导有利于学生模型的学习，而教师神经网络模型学得的神经元表达的知识与学生模型是有相似模式的，所以可以直接利用教师模型提取的特征表达指导学生模型训练。近些年，关于中间层特征的知识蒸馏方法已经有很多，出现很多较为经典的高效方法[11,12,85-88]。基于中间层知识的蒸馏方法在实践中通常需要考虑教师和学生模型的网络结构，可以将其分为同构蒸馏和异构蒸馏两种情况。

同构蒸馏指的是教师和学生模型的架构相似或属于同一系列的、层与层(Layer-to-Layer) 或块与块(Block-to-Block) 之间一一对应，如下图 4(a)所示。常见的同构蒸馏架构如 ResNet 系列、DenseNet 系列等。这是一种比较容易实现的场景，因为不需要考虑特征图大小不匹配的情况，常见的方法有

AT[12]、SP[85]、PKT[86]、DFA[87]。其中，Zagoruyko等人[12]通过将教师模型中间层的注意力图作为知识转移给学生模型，希望学生模型关注教师模型所关注的区域。SP[85]中实现了相似性保留的知识蒸馏，在预训练的教师模型和学生模型中间提取出成对样本特征并通过激活来生成相似度矩阵，而后以教师模型的相似矩阵作为学生模型的优化目标。PKT[86]是将教师模型中的特征建模为特征的分布作为知识，通过最小化差异来实现分布的拟合实现了概率知识转移。DFA[87]的思路是通过对单个教师模型不同层之间的通道特征进行聚合来模拟多个教师网络的层以减少多个教师的推理开销，通过搜索最合适的特征加权聚合增强知识迁移效果。以上方法可以看出，同构蒸馏更多的关注知识形式的构建和表达，而不过度关注模型的结构，因而相对简单明了。

异构蒸馏指的是教师和学生模型的网络结构不完全相同、难以实现层间特征图匹配的情况，结构示意如图 5(b)。这种场景下需要对中间层做相应的处理使其可用于同构蒸馏，常见的有 Hints[11]、VID[89]、AB[90]、Knowledge Flow[91]等。其中 Romero等人[11]设计了浅层的 FitNets 网络用于适配学生模型的中间层特征图大小，通过教师网络的中间层的暗示(hints)来引导学生模型向教师模型学习。Ahn等人[89]提出的变分信息蒸馏框架 VID，以最大限度地利用教师和学生网络之间特征或标签的互信息为目的，利用类似于 FitNets 的适配器架构使其可以用于任何教师-学生模型之间。Heo 等人[90]提出的激活边界 AB 框架通过对隐藏层神经元形成的激活边界进行蒸馏并实现知识转移。诸如此类方法主要解决了教师学生模型不匹配的问题，为不同结构的模型之间蒸馏提供了可能。



(a) 同构知识蒸馏 (b) 异构知识蒸馏

图 5 同构-异构蒸馏知识迁移结构图，同构知识蒸馏(a)中教师和学习模型具有相同的架构，层与层，块与块之间对应，可直接蒸馏；异构知识蒸馏(b)中教师模型和学生模型各个层或块不能完全对应需要通过桥接模块来实现蒸馏。

中间层知识蒸馏相比标签知识蒸馏更加丰富，大大提高了传输知识的表征能力和信息量，有效提升了蒸馏训练效果。并且，从中间层特征图中可以提取的知识类型更加丰富灵活，也更加具备可选择性，演化出了更多的蒸馏方法或框架。然而，很多中间层蒸馏方法不像标签知识蒸馏可以应用于各类网络间的蒸馏训练，原因在于不同架构的教师-学生模型的中间层知识表征空间通常难以直接匹配。同构蒸馏通常可以带来更好更稳定的效果，而异构蒸馏可能会因表征空间不匹配而对学生模型的训练造成困难。诸如语义分割、目标检测等高层视觉任务，使用中间层知识蒸馏可以提取注意力或上下文信息以及特征关联，比标签知识蒸馏更有优势。

目前的中间层知识蒸馏方法虽然十分丰富，也广泛应用于各类视觉、语言、推荐任务中，并且考虑了特殊场景下的特定知识，但是缺乏对多种中间层知识的整合，使它们传递不同角度的描述信息，达到互补效果。另外，如何参考模型剪枝策略对同构或异构场景下的中间层知识进行有选择地蒸馏也是值得思考的方向，如何聚合来自多个教师模型的中间层知识也是比较具有挑战性的课题。

3.3 参数知识

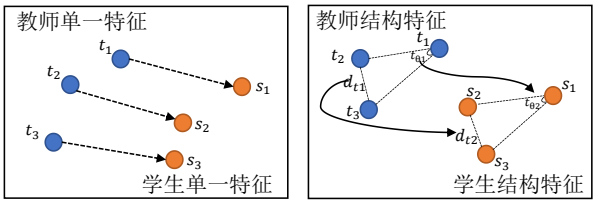
不同于中间层的特征知识，参数知识是指直接利用教师模型的部分训练好的参数或网络模块参与蒸馏训练，它通常无法作为一个独立的方法，而是与其它蒸馏方法结合使用。目前，存在两种形式的参数知识蒸馏方法：教师平均(Mean Teacher)[92-94]和模块注入(Module Injection)[31,91,95,96]。“教师平均法”作为一种稳定训练过程、保留历史信息的方式，是利用加权教师模型上一阶段更新后的参数参与下一阶段的参数更新。显然，这种方法适合于教师模型在线学习或互学习的场景，并不能用于一般的蒸馏框架，也很难用于体型和架构相差较大的教师-学生模型蒸馏训练。文献[92-94]均采用教师平均稳定半监督训练过程，保留模型参数的历史知识。

“模块注入法”同样是直接利用教师模型的参数，将教师的部分网络模块嵌入到学生模型中参与蒸馏训练，这样可以使得学生模型的模块前后处于一种接近教师模型内部的训练环境中。最早的模块注入蒸馏称为知识流(Knowledge Flow)[91]，它直接将教师模型某些模块或层的特征输出累加到学生模型的某些层的输入中，随着训练的进行教师模型逐渐脱离指导最终让学生模型独立学习。Li 等人[31]设计了一种简单而巧妙的方法用于少样本的蒸

馏训练，他们直接将训练好的教师模型的部分层取出来组装成一个新的学生模型，每两层之间通过 1×1 的卷积层匹配特征通道，经过少量样本训练就可快速恢复性能。类似地，还有一些方法通过学生模型与教师模型的模块置换与交互训练[95][96]达到辅助蒸馏的目的。模块注入的方法可以很好地为学生模型的模块训练创造一个接近于教师模型的训练环境，但是在实践操作中比较难实现，也不利于端到端的模型训练和部署。

综上所述，这两种利用参数知识蒸馏训练的方法设计都比较精巧，但对于使用场景要求相对苛刻。教师平均法要求至少两个模型是完全一样的并且同时参与训练，而模块注入法要求学生模型与教师模型的部分模块或架构具有相似性且可以置换。目前，利用参数知识的蒸馏方法刚刚兴起，在教师参数的加权调度策略、置换模块的选取方式上还有待进一步研究，可以结合诸如强化学习、元学习等自动化策略。从某种程度上讲，模块注入法可以视为模型剪枝的反向操作，如果对模型剪枝有深入了解的话会发现这两者之间有很多可以相互借鉴的之处。

3.4 结构化知识



(a) 传统的知识蒸馏　　(b) 结构化知识蒸馏

图 6 传统知识特征与结构化知识特征对比，传统的知识蒸馏(a)主要是在特征上直接蒸馏；结构化知识蒸馏(b)在特征之上构建特征之间的结构关系（如：距离和角度[82]）。

传统的知识蒸馏方法默认样本或特征之间相互独立，采用点对点的知识迁移方式；而结构化知识蒸馏则关注到样本之间的关系（类间关系）或者样本特征内部的上下文关系（类内关系）等结构化表征，两者的对比如图 6 所示。结构化知识的构建对知识的具体位置没有严格的要求，可以是模型输出层的类别，也可以是中间层特征。模型对样本之间的差异度量建立在结构化知识之上。而对于同一种结构化知识的差异度量方式通常可以有多种选择。采用样本间关系度量的结构化知识蒸馏可以很好地突破同构、异构蒸馏的限制，所得到的知识矩阵只和批训练中的样本数量有关，和模型的中间层特征图通道、大小等属性没有关联。而关注特征上下文关系的结构化知识蒸馏通常要面对异构蒸馏中特征图难以匹配的问题。

对结构化知识蒸馏影响最重要的是 2019 年发表在 CVPR 上的 RKD[97]和 IRG[98]方法，它们都是关注样本间关系的蒸馏方式。RKD[97]通过建立样本间的角度和距离双重关系度量，使得细粒度图像分类任务上的学生模型可以关注到教师模型所提取到的各类别样本（包括类内的、类间的）之间的结构化关联，解决了度量学习中的细粒度知识迁移难题。类似地，IRG[98]同样是提取样本之间的结构化关联，但是它采用了计算实例在神经网络层与层之间的关系图的方式。RKD 只利用模型输出的 embedding 作为样本表示，而 IRG 提取的实例关系图与模型结构密切相关。因此 RKD 更加通用和简便，但 IRG 只能用于同构蒸馏，而且计算量更大。Chen 等人[99]提出了一种更加简便的结构化蒸馏方式，它显式地建立了类别内部和类别之间样本的结构化表示，这与 RKD 利用随机采样的方式建立的未明确样本所属类别结构化关系不同。

对于提取特征内部上下文关系等作为结构化知识的蒸馏方式与中间层知识蒸馏的界限划分并不是十分的明确，这里只介绍几种明确了结构化概念的蒸馏方法。Liu 等人[78]针对语义分割任务提出了一种利用像素以及像素之间相似性的结构化蒸馏方法，这可以帮助学生模型捕获语义特征的长距离依赖关系。You 等人[100]提出区域间亲和度蒸馏的方法用于车道线检测，将给定的道路场景图像分解为不同的区域，并将每个区域表示为图中的一个节点，然后根据节点在特征分布上的相似性，建立节点之间的成对关系，形成区域间亲和力图。Tao 等人[101]通过逐步建立样本之间的结构化图表示，实现了一种新颖的少样本增量学习方式，充分利用了学得的样本关联信息作为未知样本推理的依据。Li 等人[102]借鉴了语义分割中用到的特征通道注意力和空间注意力机制，以此建立图像中的语义结构关联，实现了一种新的图到图翻译网络的蒸馏方法。

结构化知识挖掘到了教师模型提取的样本特征之间的关联，相对于单样本内的独立知识，它显式地建立了所提取特征中携带的额外信息，这使得对大型教师模型知识的利用更加彻底和充分，为知识蒸馏打开了新的视角。结构化知识蒸馏还停留在基础阶段，像在线学习、互学习、多模型学习等方

向还缺少与结构化知识的深度结合，在人体姿态估

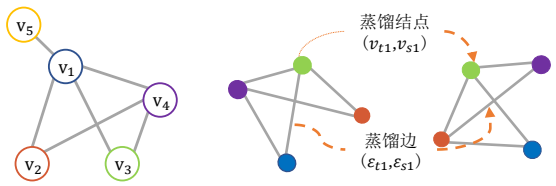
表 2 不同知识形式的代表性蒸馏方法在 CIFAR100 数据集上实验结果

知识形式	代表方法	ResNet-56 ResNet-20	ResNet-110 ResNet-20	ResNet-110 ResNet-32	VGG-13 VGG-8
	教师表现	**72.34**	**74.31**	**74.31**	**74.64**
	学生表现	**69.06**	**69.06**	**71.14**	**70.36**
标签知识	KD[10]	70.66	70.67	73.08	72.98
中间层知识	AT[12]	70.55	70.22	72.31	71.43
	SP[85]	69.67	70.04	72.69	72.68
	PKT[86]	70.34	70.25	72.61	72.88
	VID[89]	70.38	70.16	72.61	71.23
	AB[90]	69.47	69.53	70.98	70.94
	FitNets[11]	69.21	68.99	71.06	71.02
结构化知识	RKD[97]	69.61	69.25	71.82	71.48

计、语义分割、多目标检测等高级视觉任务以及自然语言处理、推荐系统等领域的应用也十分需要对结构化信息的提取和迁移。另外，对于序列或视频相关的任务，序列节点之间的前后关联也可以作为一种特殊的结构化信息用于知识蒸馏。总之，关于结构化知识蒸馏可以深入研究和挖掘的点非常多，是一个非常活跃的方向。

3.4 图表示知识

图表示知识可以看成是结构化知识的进一步拓展和特殊形式。图表示学习逐渐成为研究热点，主要原因在于现实生活中很多场景中的数据形式是非结构化的(如分子模型生成、3D 点云建模)，需要用图(Graph)结构来建模，它可以表示节点之间更复杂的多种关系。而传统的以自然语言为代表的时序数据和以图像为代表的表格数据在表征内部特征之间的关系上相对较弱。因此，可以通过一定方式将其转化为图表示的结构以强化特征之间内在的联系。与之相应的知识蒸馏方法也应运而生。



(a) 图表示知识 (b) 图表示知识蒸馏

图 7 图表示知识与图知识蒸馏示意图，图表示知识(a)通常构建成节点和边的连接形式，知识蒸馏需要建立在边表示的节点关系或局部图结构上(b)。

目前，使用图表示知识的蒸馏方法主要集中于两类场景：一是从经典深度神经网络中提取特征的图结构化关系表示知识[103-106]，这与前面所介绍的四种知识蒸馏形式都属于经典深度神经网络的蒸馏范畴；二是最近非常热门的图神经网络(Graph Neural Networks, GNN) 上的知识蒸馏[107-112]，这就完全依托图卷积网络模型中的节点特征以及它们之间的关联。如 Lee 等人[103]认为传统的 KD 方法很少定义数据的内部关系，也无法生成嵌入知识。因此，提出了利用注意力网络从教师网络中提取知识。在知识嵌入的过程中利用多头注意力(Multi-head Attention) 方法将教师网络提取成图，并通过多任务学习将关联诱导偏差传递给学生网络。Zhang 等人[105]通过模型输出的软标签和中间层特征构造图表示来从多个自监督任务中转移知识，前者蒸馏软标签知识解决多元分布联合匹配问题，而后者从成对集成表示中蒸馏内部关联来解决不同特征之间的异构性挑战；该方法可以显著减少从教师中提取的知识的冗余信息。Lassance 等人[106]通过网络中间层的特征图来捕捉潜在空间的几何结构，并在训练过程中引入各自邻接矩阵之间的差异度量，利用从教师网络提取的图信息训练学生。

以上几种方法，都是针对欧氏空间中常见的图像处理问题，可以通过不同的方式在模型内部将其构建为具有非富信息的图表示结构，这种方式能够压缩模型同时提高模型的收益。而 Ma 等人[104]提出的方法是针对图结构数据本身，在 DeepGraph 网络中利用 HKS 可以将同构图映射为具有丰富信

息的同构表示，而后构建辅助任务实现多任务知识蒸馏。这种多任务学习方法能够提高原始学习任务

表 3 不同“知识”表达形式的优缺点

知识形式	优缺点	解决的问题	适用场景
标签知识	**优点**：方法简单通用，易于实现。 **缺点**：知识单一，依赖于损失函数的设计且对参数敏感。	较为通用，一般作为各种任务中知识蒸馏的基础。	适合分类，识别，分割等几乎所有任务。
中间层知识	**优点**:具有丰富知识信息,能够满足复杂任务对知识的需求。 **缺点**：知识形式多样，无法有效的整合且互相影响，试错代价高。	解决了输出层知识信息单一，不够丰富的不足，能够为模型提供较为丰富的特征知识。	适用于安全隐私要求相对不高，教师模型可访问的场景;对特征依赖较大场景且模型准确率有较高要求的情况。
参数知识	**优点**：模型训练稳定且效率较高，设计精巧。 **缺点**：较难实现，不利于端到端的模型训练和部署。	解决学生模型训练相对较慢，无法快速适应并且训练不稳定的问题。	适用于在线学习、互学习等较为苛刻场景，需要教师和学生模型具有较为类似的架构。
结构化知识	**优点**：对特征之间和特征内部关系的表征较强。 **缺点**：计算开销较大，优化成本相对较高。	单一样本表征能力不足，信息较为简单且无法满足复杂任务的要求。	适用于复杂任务中的关系度量且对上下文信息具有较高要求的任务中，如细粒度分类等。
图表示知识	**优点**：丰富了知识的表示形式且能够有效提高学习任务的性能。 **缺点**：特征构建较为复杂，计算开销较高，泛化性能较差。	针对非结构化知识表示的问题，复杂的节点关系信息表示。	适用于非结构化的数据，如 3D 点云，分子式分类等；结构化数据在通过一定转化后也可适用。

的预测性能，特别是在训练数据较少的情况下。图神经网络的知识蒸馏，是一个初露头角的研究方向。由于图神经网络在计算机视觉、自然语言处理、推荐系统等任务上的大量运用，学者们开始逐渐关注它的迁移训练和压缩等问题。Yang 等人[112]首次尝试了图神经网络知识蒸馏的研究，为了能够将知识从教师图卷积神经网络(Graph Convolutional Networks, GCN) 模型传递给学生模型，提出了一个局部结构保留模块，该模块显式地说明了教师模型的图卷积层节点拓扑语义，并将其传递到学生模型的局部表示。类似地，Zhang 等人[107]提出了一种稳定的图卷积蒸馏框架，通过度量教师和学生模型中节点和边的稳定性，达到迁移稳定知识的目的，有助于提升蒸馏效果。最近，图神经网络的知识蒸馏还演化出了互学习蒸馏[108]、自蒸馏[109,110] 等新的形式。另外，动态图表示学习策略基于不同的图神经网络结构来捕获图随时间的演变过程，训练参数量大且推理过程耗时。因此，Antaris 等人[111]提出了 Distill2Vec 的蒸馏策略，用于训练具有少量可训练参数的压缩模型，从而减少在线推理的时延并保持较高的预测准确性。

目前，随着图神经网络的研究逐渐兴起，关于图知识蒸馏的研究也逐渐被研究人员所关注，其存在的主要挑战在于图结构数据表示局限于特定的结构化的数据和特定的类型，因此泛化性相对较差。其次是图结构空间距离的度量是一个挑战。由于大部分蒸馏方法源于图像任务和经典卷积神经网络，因此图神经网络上的知识蒸馏方法并不成熟，还有非常多问题亟待研究。关于图结构信息在图像任务上的应用非常受限，相应的知识蒸馏方法也有待开发。

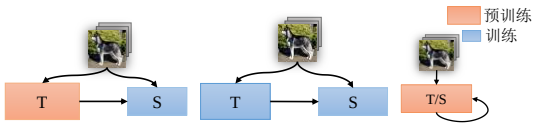
4 学习方式

类似于人类教师和学生间的学习模式，神经网络的知识蒸馏学习方式也有着多种模式。其中，学

生模型基于预训练好的、参数固定的教师模型进行蒸馏学习被称为离线蒸馏(Offline Distillation)。相应地，教师和学生模型同时参与训练和参数更新的模式则称为在线蒸馏(Online Distillation)。如果学生模型不依赖于外在模型而是利用自身信息进行蒸馏学习，则被称为自蒸馏学习(Self-distillation)，如下图 7 所示。一般而言，蒸馏框架都是由一个教师模型和一个学生模型组成，而有多个模型参与的蒸馏称为多模型蒸馏(Multi-model Distillation)；目前，大部分蒸馏框架都是默认源训练数据集可用的，但最近的很多研究在不使用任何已知数据集的情况下实现蒸馏，这类统称为零样本蒸馏（又称为无数据蒸馏，Zero-shot/Data-free Distillation）。特别地，出于一些隐私保护等目的，教师模型可以享有一些特权信息而学生模型无法访问，在这种约束下，形成特权蒸馏(Privileged Distillation) 学习。接下来，将分别介绍不同蒸馏学习方式的代表性工作。

4.1 离线蒸馏

离线蒸馏通常是在拥有预训练完备的高性能教师模型的前提下进行的。在蒸馏学习过程中教师模型只进行推理而不更新参数，学生模型每个训练周期都从教师模型获得固定不变的知识，如图 8(a)所示。相对于在线蒸馏而言，它的主要优点在于，可以灵活选择一些预训练好的大型模型作教师，在蒸馏过程中大型教师不需要参数更新，而只需要关注学生模型的学习，这使得训练过程的部署简单可控，大大减少了资源消耗和训练成本。



(a) 离线蒸馏 (b) 在线蒸馏 (c) 自蒸馏

图 8 学习方式分类结构示意图（T 为教师模型,S 为学生模型，下同）。

然而，离线蒸馏通常不能满足一些多任务、多领域学习场景的需求，其原因在于多个任务/领域模型都需要参与训练过程并且从其余任务/领域中学习有利知识，这就要求多个模型必须同时在线参与蒸馏学习。离线蒸馏不能保证教师模型与学生模型的学习过程相匹配，也不能根据学生模型的学习状态实时调整教师模型的知识提炼过程，如果训练完备的教师模型和学生模型的预测性能差距很大，则会影响学生模型在初始阶段的学习[48]。但是，目前离线蒸馏仍然是主流的蒸馏学习范式。

4.2 在线蒸馏

在线蒸馏过程中，教师模型与学生模型都会同步更新参数，所传输的知识在每个阶段也都会不断更新（图 8(b)所示），这不同于离线蒸馏中预训练完备的教师模型在蒸馏阶段只进行推理，每个阶段输出固定不变的知识。特别地，在线蒸馏不一定完全遵循教师-学生的训练框架，其学习形式也具有多样性。但是，共同的特点是所有参与蒸馏的模型是可学习的，典型的模式有互学习、共享学习和协同学习。

互学习：互学习(Mutual Learning) 的特点是将两个或多个学生模型一起训练并将他们的输出知识作为互相之间的学习目标。如 Zhang 等人[113]提出两个学生模型之间互相学习，共同提高学习效果，并且还扩展到了多个学生模型互学习的场景。Chen 等人[114]通过两级蒸馏实现多样的同伴来共同学习，在一级蒸馏过程中提取各个模型的高维特征和 logits，通过注意力机制将特征和 logits 按照权重提取并传递给各个学生，在第二级蒸馏将各个学生的知识传递给组长，并用作最后的模型部署。此外，笔者在文献[115]提出深度对抗互学习的特征提取方法从源域中学习情感分类来提升自适应目标域的情感分类。该方法也是属于互学习。互学习的优势在于模型之间可以相互促进实现互补。

共享学习：不同于互学习，共享学习在多个训练模型中需要通过构建教师模型来收集和汇总知识，并将知识反馈给各个模型，以达到知识共享的目的。如 Lan 等人[116]提出通过添加多个分支来重新配置网络，这些分支和目标网络共享低级层。每个分支构成一个单独的模型，它们的集合用于构建教师模型。教师模型会在训练过程中实时的收集知识，然后这些知识被提取并反馈给各个分支，以闭环的形式加强模型学习。在应用中可以根据部署的需要丢弃或者保留辅助分支。另外，Xie 等人[117]也提出了类似的架构，通过简易的网络来实现卷积而后在线蒸馏的过程中将多个网络并联起来训练提起特征，并将特征融合用于构建一个强大教师模型。在训练过程中，通过相互的学习来提高师生的预测能力。这种多分支共享的蒸馏学习模式除了用于提升多分类器效果，还可以被用于多任务学习[54,118]、长尾数据训练[119]等。

协同学习：协同学习(Collaborative Learning)是近些年刚刚兴起的一个研究主题。其概念类似于互学习，协同学习主要是在任务上训练多个独立的

分支后实现知识集成与迁移并实现学生的同时更新。Song 等人[120]引入了协同学习，在相同的训练数据上同时训练多个分类器，以提高泛化和对标签噪声的鲁棒性，而不需要额外的推理代价。它从辅助训练、多任务学习和知识提炼中获得优势。Guo 等人[121]提出了一种通过协同学习的高效在线知识提炼方法，该方法能够持续提高具有不同学习能力的深度神经网络的泛化能力。Guan 等人[122]将协同学习方法应用到了 Style-GAN 上，提出了一个新的协作学习框架。该框架由一个高效的嵌入网络和一个基于优化的迭代器构成。随着训练的进行，嵌入网络对迭代器的隐藏码给出了合理的初始化。另一方面，迭代器更新的隐藏码反过来监督嵌入网络。最后，通过嵌入网络，通过一次前向传递可以有效地获得高质量的隐藏码。

在线蒸馏相比于离线蒸馏能够针对不同任务在没有预训练模型的情况下实现知识学习和蒸馏，也有助于各个模型在互相学习的过程中调整自身训练和更新贡献的知识，更好地实现优势互补，对于多任务学习等特殊场景具有很大优势。但是在线学习最大的挑战在于在线模型数量的增加虽然能够带来一定的效果提升，但是也造成计算资源的消耗。其本身对于模型压缩意义不大，更适合用于知识融合或在多模态、跨领域等场景中发挥价值。

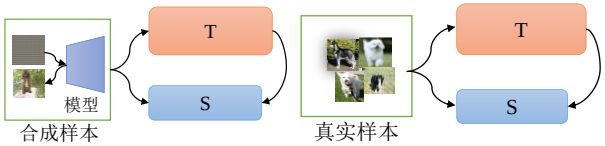
4.3 自蒸馏

自蒸馏是一种比较特殊的蒸馏模式，它是指学生模型从自身输出的知识中进行学习，通常是将深层信息回传给浅层指导训练过程，不需要其他教师模型的辅助，如上图 8(c)所示。自蒸馏目前主要作为一种提升下游任务模型性能的手段，而不以模型压缩或迁移学习为目的。

最早关于自蒸馏学习的工作[63,64,66]发表在 ICCV 2019 会议上，Zhang 等人[63]和 Phuong 等人[65]的思路基本一致，都是在卷积神经网络的中间每一层接入一个提前预测分类结果的分类器，由模型最后的主分类器输出的 logits 引导中间各层的早期预测。不同于文献[63,65]，Hou 等人[66]提出的基于自蒸馏的车道线检测算法，它是利用类似于中间层注意力图[12]，每一层接受后一层的注意力引导训练，通过将模型深层次的特征提前传给浅层学习来提升车道线检测模型的性能，这其实和文献[63,65]中提前预测分类结果的原理类似。这三项工作奠定了自蒸馏的基础，也是非常经典有效的模型提升方法。类似地，如 Yun 等人[64]提出的基于正则化的类别自蒸馏方法，主要是通过同一类别内不同样本之间的 logits 知识的相互利用来提升分类模型性能，这其实是利用了样本级的信息自蒸馏而不是模型的自蒸馏，与自监督学习有相通之处。而 Lee 等人[67]结合自监督学习中的数据增强方式，利用旋转增强的同一数据样本通过联合分类器输出的多个预测 logits 聚合后，用于指导主分类器的学习，这同样是利用了样本级的自蒸馏。最近的工作还有文献[123][124]，Huang 等人[123]提出了一种综合注意力的自蒸馏方法用于提升弱监督目标检测模型的效果；Liu 等人[124]则延续文献[63,65]中的思路并将自蒸馏应用于元学习中，该方法的创新之处在于能够为中间层生成更具兼容性的软目标，而以往自蒸馏输出层软标签应用于中间可能存在深层和浅层模块间不匹配的问题。

自蒸馏的提出主要是针对传统的两阶段蒸馏方法必须预先训练大型教师模型导致的耗时问题，而且教师模型和学生模型存在能力不匹配的问题，使得学生无法有效学习教师的表征。自蒸馏可以在没有教师模型指导的条件下达到学生模型性能的自我提升，但是在同样的任务和实验环境下，究竟自蒸馏和教师-学生框架下的经典蒸馏方法孰优孰劣目前还缺少相关实验对比和考证。另外，关于自蒸馏原理的分析工作[125,126]主要是从正则化的角度对其进行解释，这对于探索新的自蒸馏方法具有重要指导意义。

4.4 无数据蒸馏



(a) 有数据蒸馏　　(b) 无数据蒸馏

图 9 传统知识蒸馏模型和无数据知识蒸馏的结构对比图。无数据蒸馏需要通过噪声合成等效样本同时将知识传递给学生模型。

无数据蒸馏(Data-free Distillation) 的概念（如上图 9(b)所示） 是由 Lopes 等人[127]在 2017 年提出，在后来的一些工作中[25,80,128,129]也常被称为零样本蒸馏(Zero-shot Distillation)。它指的是给定一个预训练完备的教师模型，在没有任何源训练数据集的情况下进行知识蒸馏并训练学生模型，使之达到与有数据监督训练接近的表现。由于无数据蒸馏的要求较为苛刻，这使得它成为当前知识蒸馏研究方向最大的挑战之一，直到最近两年才涌现出大量相关

成果。因为在没有源训练数据的情况下，能够利用的信息只有教师模型中保存的参数。这就需要利用某些方法和一些先验信息从中逐步提取相关信息甚至还原部分样本用于训练学生模型。其中最大的困难就在于准确还原样本，使之尽可能与源训练数据的样本特征接近。而这些样本并不完全与源数据一致，因此也被称为等效样本或者伪样本。目前的无数据蒸馏方法根据还原等效样本的方法可以分成两类：对抗合成方法[80,129,130]和噪声优化方法[127,128,131,132]。这两种方式都要依据教师模型参数中保留的先验信息来合成等效样本，不同之处在于对抗合成方法利用生成器与教师模型形成对抗学习来估计样本分布，而噪声优化则是直接优化每次输入教师模型的噪声变量直到收敛。至于教师模型中可以利用的先验信息，一般有激活层输出的激活信息[127,130,132]、批量归一化层的统计量(Batch Normalization Statistics, BNS)[131]、输出层的信息熵[130,131]以及教师-学生之间的差异度量[80,129]。

最初，Lopes 等人[127]认为教师模型训练后的激活层中输出的信息表征了网络原训练数据的敏感性，因此将其作为元数据重构训练样本，用于对学生模型进行训练。之后，在 2019~2020 年涌现出大量新的更有效的无数据蒸馏方法，Bhardwaj 等人[132]通过 10% 的真实 CIFAR-10 图像输入教师模型产生的激活向量构建元数据，以此来合成为样本用于模型压缩；由于其使用了元数据，因此研究人员普遍认为这并不是严格意义上的无数据蒸馏。Nayak 等人[131]在不使用任何元数据的情况下，同样基于激活层信息，利用样本间的关系，构建狄利克雷分布来从噪声中还原等效样本。而 Chen 等人[130]提出的 DAFL 对后面的工作产生了重要影响，也是使无数据蒸馏重获关注的开创性工作。DAFL[130]方法虽然与后面的方法相比并不高效且存在一定的局限性，但是它所提出的利用一个额外的生成器与教师模型形成对抗学习的框架十分重要，另外还利用了 3 种教师模型中的先验信息作为鉴别合成数据的依据，包括激活信息、输出层信息熵和类别信息。最近，Yin 等人[128]创造性地利用目前深度学习框架中常用的批量归一化(Batch Normalization, BN) 层组件中保留的均值和方差作为先验信息，提出了 ADI 方法，通过优化噪声可以得到在视觉效果上接近真实样本的图像。对比利用 BN 层统计量(BNS) 和激活层信息的方式，BNS 实际上是把每层的输出看作一个高斯先验分布，具有更强的可解释性，但是其要求输入的噪声图像的 batch size 要足够大，才能利用 BNS 的优势；如果 batch size 过小，则无法准确拟合先验的高斯分布。ZSKD[129]和 DFAD[80] 都考虑到利用教师与学生模型之间预测输出的差异度量，通过生成器最大化其差异来合成等效样本，反过来通过最小化学生与教师的差异进行知识迁移。不同的是，ZSKD[129]利用 KL 散度(Kullback-Leibler divergence) 来度量差异，而 DFAD[80]提出 KL 散度不可避免地会使得优化过程陷入一些异常的伪样本而难以进行，因此使用L_1度量更加有效。而通过实验结果也可以看出，DFAD 更加高效和通用，其不受网络输出的限制，可以用于图像分类、语义分割等各种任务。

无数据知识蒸馏已经显示出巨大的潜力，能有效克服实际场景中数据不足、隐私保护等挑战，为模型压缩、迁移学习等商业应用的部署提供了可能。但是，目前的难点在于模型很难恢复质量较高的图像，主要表现为分辨率低、与原数据集差异大、噪声影响严重、数据多样性不足、依赖于标签信息、局限于分类任务。因而，在一些对数据要求较高的密集型预测任务（如语义分割）中，还无法满足要求。另外，无数据蒸馏的训练框架相比有数据驱动的蒸馏框架十分耗时，尤其是巨大的样本量和迭代次数，这给还原高分辨率的图像带来了阻碍。如何降低数据还原的巨大成本、提升无数据蒸馏的知识迁移效率是该领域亟待解决的问题，也是最有潜力的研究方向。完全不依靠数据实现伪样本的还原并不可靠，寻找如何探索和利用(Exploration and Exploitation)教师-学生之间特征度量空间的差异以提高复杂样本的分类和还原效果是值得思考和探究的。

4.5 多模型蒸馏

由于传统的教师-学生蒸馏框架下都是一个教师指导一个学生的训练，这可能存在单个教师输出的知识并不完全可靠的问题。因此就有研究者提出了多教师蒸馏、集成学习的多模型蒸馏方式。多教师蒸馏是从多个预训练完备的教师模型中提炼知识用于指导学生模型；而集成学习则是一种在线蒸馏或者互学习的方式，多个学生模型同时参与训练，各自集成其他模型输出的知识后进行学习。

4.5.1 多教师蒸馏

多教师蒸馏的研究重点在于设计合适的知识组合策略用于指导学生，学习多个教师的优点而摒弃不足。由于教师模型网络架构的多样性和复杂

性，多教师蒸馏主要关注于分类问题，尤其是教师权重的分配与组合，因此相对局限，多模型蒸馏的相关研究目前也主要集中在集成学习上。

You 等人[56]在 2017 年提出了多教师蒸馏的框架，将多个教师模型输出的 logits 软标签进行平均后提供给学生模型学习，他们还引入了 triplet 损失函数，利用多教师中间层特征的相对差异指导学生的特征学习。后来，笔者[57]为了纠正平均分配教师权重的不合理性，在训练中体现不同教师对不同样本知识的贡献，引入了自适应的多教师多层次蒸馏方法，根据每个样本的潜在表征自动选择分配合适的教师 logits 权重。Wu 等人[60]为了提升行人重识别系统在目标领域上的稳定性，根据验证集的预测风险分数为多教师分配合适的权重来聚合知识。Son 等人[58]设计了一种利用多个不同深度的辅助教师密集引导的蒸馏方法，教师按深度由深到浅分为多个等级，每一级同时接受前面所有教师的软标签指导，直到最小的学生模型，以此最大程度减少层级间的学习差距。类似地，Shi 等人[59]则在人脸识别模型上用了另一种直接拼接多个教师的 logits 然后进行 PCA 降维的方式。Shin 等人[61]将多教师-单学生的蒸馏架构扩展到目标的视觉多属性识别任务，每个教师专门学习一种属性，然后综合多教师的知识传递给学习，实现学生的多属性识别学习，类似于一种多任务的学习模式。从上述多教师蒸馏方法的对比可以看出，核心的设计在于多个教师软标签知识的组合策略。也许这对于基于软标签的学习任务是有效的，但是目前多教师蒸馏的局限也在于此。由于大部分任务是采用现成的与训练完备的教师模型，各个教师的结构可能不一致，导致中间层特征知识难以有效组合和增强，而通常中间层知识蒸馏更为高效。另外，从文献[61][119]的思路可以受到很多启发，多教师蒸馏对于多任务、多模态学习等有很重要的指导意义，可以解决传统端到端训练方式面临的许多困难。

4.5.2 集成学习

集成学习类似于多教师蒸馏，关键在于多个模型的知识集成策略的设计，使其达到优势互补的效果。不同的是，集成学习没有严格意义上的教师模型参与，所有学生模型都同时学习和更新参数。并且，它通常采用多个完全同构的模型，因此对中间层特征的利用度很高。此外，集成学习还要求每个学生将自己的知识汇总到一个集成器中，而后集成器再将汇总知识分发给每一个学生。

Lan 等人[116]在 2018 年提出的 ONE 方法是深度神经网络集成学习的先驱，其实 ONE 与 DML[113]的方法异曲同工，只不过 ONE 采用了多分支集成学习的方式，通过门控决策不同分支的 logits 权重。Chen 等人[114]也采用的相似的多分支或者多同伴的方式，额外引入了中间层特征的集成。Park 等人[133]借鉴 AT[12]方法，拓展出了基于注意力特征的集成学习方法，显然这对于各个子模型的架构要求十分苛刻。Walawalkar 等人[134]借鉴 FitNets[11]，在多个异构学生模型间引入特征适配网络解决了集成学习中这一问题，但是其创新性有限。最近，Du 等人[135]提出了一种新颖的基于梯度的“求同存异”的软标签与中间层特征集成学习策略，具体地，根据每个子模型计算的梯度为每个子模型输出的知识分配合适的权重来控制集成。

从实践的角度讲，巧妙设计各种知识集成策略用于蒸馏并不一定比选取一个合适的教师模型和有利的知识传递形式更加方便高效，而且在经典教师-学生蒸馏框架下有非常多优秀的预训练教师模型可供选择，单个教师所输出的知识足以胜任学生的指导任务；而集成学习的框架部署相对复杂许多，多模型同时在线训练无疑会造成严重的计算和存储负担，而获得的性能收益不会因此线性增长。因此，从宏观和更长远的角度来看，在单教师-学生蒸馏框架下研究更加高效的、准确的、通用的知识提炼和传输方法更有意义，也可以为模型压缩、迁移学习等提供更简易的解决方案。相比多教师蒸馏，集成学习尚未在多任务、多模态等设定下发挥优势。

4.6 特权蒸馏

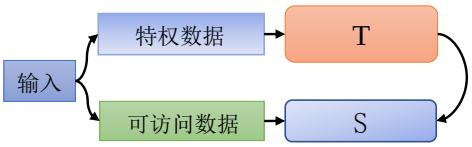


图 10 特权蒸馏结构。特权数据只能教师模型访问，学生模型无法直接访问，学生模型需要通过教师模型来学习。

特权蒸馏是一种与架构或者蒸馏模式无关的场景，它是针对参与训练的数据而言的。“特权”指的是教师模型在训练过程中可以访问一些特定信息，而学生模型在训练中无法直接享有这种特权，只能通过从教师模型蒸馏学习获得，如图 10 所示。特权信息学习最早是由 Vapnik 等人[136]提出的学习范式，他们在传统的机器学习模型支持向量机(SVM)

上利用了特权信息，并通过严格的公式推导了学习　　的过程，描述了利用特权信息显著提高学生学习速

表 4 不同蒸馏方法的优缺点比较

学习方式	优缺点	解决的问题	适用场景
离线蒸馏	**优点**：灵活可控，易于操作，成本较低。 **缺点**：无法满足多任务、多领域任务。	通用的基础学习方法，主要针对单任务学习。	适用于单任务学习，安全隐私要求相对不高，教师模型可访问的场景；
在线蒸馏	**优点**：参与训练，实现模型之间互补。 **缺点**：操作较为复杂，模型开销较大。	在没有预训练的情况下实现知识学习和蒸馏。	适用于多模型、多任务、跨领域等场景。
自蒸馏	**优点**：结构简单，开销较小，训练稳定。 **缺点**：缺少理论支撑。	解决模型过拟合和蒸馏开销较大的问题。	适用于并行化训练，低开销且对模型准确率有较高要求的场景。
无数据蒸馏	**优点**：能够有效保护数据安全隐私。 **缺点**：复杂场景下的任务准确率不高。	解决模型部署开销大、数据安全隐私等问题。	适用于对数据安全隐私要求较高的场景。
多模型蒸馏	**优点**：能够利用的特征较为丰富，泛化性能较高。 **缺点**：特征构建较为复杂，计算开销较高。	单个教师模型输出知识不可靠。	对模型准确率和泛化性能要求较高而对模型开销相对较低的场景。
特权蒸馏	**优点**：提升学习效果，降低训练难度，提高数据保护。 **缺点**：教师输出信息仍有一定安全隐患。	解决学生模型无法访问数据的问题。	应用数据隐私较高的场景。

度的两种机制。Tang 等人[137]提出的特权蒸馏方法，通过无监督学习的外部数据源来构建特权信息，允许其被保留在一个多任务学习设置中，使用一种新的特征匹配算法从原始特征空间和特权信息空间中抽取样本，并将样本相似度信息映射到联合潜在空间中。Wang 等人[138]提出了名为 KDGAN 的教师模型、分类器和鉴别器组成的三模型特权蒸馏框架，教师与鉴别器享有完整数据作为特权信息，而所训练的分类器只能访问部分离散数据，鉴别器用于鉴别目标分类器预测标签的真伪，教师模型向分类器传输软标签知识。Bhardwaj 等人[139]和 Zhang 等人[140]设计了一种简洁而实际的特权蒸馏场景，其主要目的不是用于隐私保护，而是减少学生模型的训练数据量。前者[139]是用于视频分类任务，而后者[140]是人体姿态估计任务，他们的共同点就是教师模型训练中使用完整的视频帧，而学生模型的训练只接受少部分视频帧，利用知识蒸馏由教师向学生传达额外的信息。

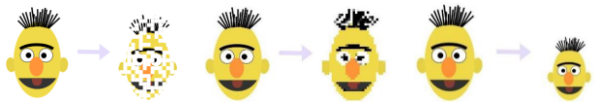
特权蒸馏目前主要用于一些隐私保护的场景，在知识传递形式上主要是以软标签信息为主，学习形式没有严格约束。相比于无数据蒸馏，特权蒸馏可以让教师模型利用特权信息而学生模型可以间接地通过蒸馏学习获得这些信息，这无疑可以提升学生的学习效果，而且降低了训练难度。虽然大部分特权蒸馏方法都要求学生无权访问特权信息，但是教师模型输出的特征中保留了大量可以还原特权数据的信息，严格上讲，这同样也会有一定的隐私泄露的风险。

5 学习目的

5.1 模型压缩

模型压缩[141]是知识蒸馏提出的最初目的，也是目前最主要的应用方向之一。深度神经网络是人工智能的各个领域快速发展的基础，但是随着任务的复杂性增加、性能要求愈高，导致神经网络模型的结构愈加复杂，这直接导致了计算成本的急剧上升，严重限制了其在移动嵌入式设备上的部署和应用。因此，模型压缩逐渐受到学术界和工业界的广泛关注，各类压缩方法，如剪枝[125,142]、量化[143-146]、低秩分解[147,148]、高效结构设计[149-151]以及知识蒸馏等。图 11 形象地展示了三种主要的模型压缩方法的原理。剪枝（图 11(a)）是除去网络每一层中贡献最小的一部分参数，可以是非结构化的一些参数，也可以是整个卷积通道，甚至整个网络层。这通常会破坏网络的结构，而且需要迭代地训练和搜索最优的剪枝结构。量化（图 11(b)）是从网络参数表达上，将冗长的浮点数截断至低比特（通常是 8-bit 以

下）表达的整型数，这完整保留了模型的宏观架构，而且可以实现较高的压缩率。蒸馏（图 11(c)）是直接从大容量的教师模型向小容量的学生模型迁移知识来提升性能。这不需要改变网络结构或内部操作，实现起来比较灵活，几乎适用于任何模型。



(a) 剪枝 (b) 量化 (c) 蒸馏

图 11 三种主要模型压缩方法的原理示意图。箭头左边为原始模型，右侧为压缩模型。

上述介绍的大部分知识蒸馏方法都是以模型压缩为目的，这里不再赘述。本节将主要介绍一些蒸馏与其他模型压缩方式结合的工作，以揭示知识蒸馏在模型压缩方面的巨大潜力。

蒸馏+剪枝：Ashok 等人[152]利用强化学习对网络进行裁剪，从 Layer Removal 和 Layer Shrinkage 两个维度进行裁剪，一个是对层判断是否进行裁剪，一个是决定一层中参数的裁剪，同时利用软标签蒸馏训练剪枝模型。Gao 等人[153]和 Miles 等人[154]都是将中间层特征蒸馏方法 Hints[11]融合剪枝形成循环迭代过程，主要是不断缩小剪枝模型与原模型各层的差距。Chen 等人[155]则实现了跨域模型的协同剪枝，通过两个领域共享的掩码操作来决定需要剪枝的滤波器。最近，Luo 等人[156]将知识蒸馏与 Mixup[157]和标签增强结合，形成一种自监督机制，有效扩充了数据，在少样本情况下实现了剪枝。

蒸馏+量化：由于量化可以保持稳定的模型结构，使得量化模型与原始模型各层之间可以完全匹配，因此蒸馏与量化结合的相对容易。Zhuang 等人[158]将 FitNets 移植到量化模型的训练上，由于架构之间良好的匹配，这非常容易扩展至多层蒸馏。Kim 等人[159]实现了一种三阶段的训练策略，首先量化模型进行自学习，即常规有监督训练，然后全精度的教师模型与量化模型互学习，最后由训练完备的教师模型对量化模型蒸馏训练。Zhuang 等人[160]利用量化模型与原始全精度模型结构上的一致性，在各层之间添加一个辅助模块用于浮点精度数据到量化精度的过渡，逐层进行蒸馏训练。二值网络是极致量化的模型，其参数仅由 0、1 或者-1 与 1 表示，实现了超高的压缩率，精度也会严重损失，因此 Ye 等人[161]利用了残差学习指导恢复二值网络的性能。另外，在 4.4 节中所介绍的基于无数据蒸馏的量化方法[162-164]，是目前非常活跃的研究方向。

其他：Lin 等人[165]提出了一种整体的低秩分解压缩方法，将每个卷积层滤波器通过奇异值分解为两个新的上三角矩阵，减少了参数量，然后在中间层和输出层加持局部和整体的知识蒸馏增强训练效果。文献[166]也是低秩矩阵分解与蒸馏的结合，从每个卷积核中提取n个 1x1 的点，这n个点每个点单独构成一个和原来卷积核大小相同的矩阵$M_{1\sim n}$，再用n个可训练的权值参数对$M_{1\sim n}$重构形成一个新的卷积核，而蒸馏方法上借鉴了 FitNets[11]。

目前，知识蒸馏与其他压缩方法的结合还处于比较基础的阶段，缺少很多针对性设计，比如剪枝中不断变化的网络结构，量化参数与全精度参数的不一致表达等。特别地，基于无数据蒸馏的剪枝和量化方法刚刚兴起，是一个非常有前途的研究方向，这可以使得模型压缩的场景更加广泛，只需要一个训练好的模型而不需要额外数据或信息就可以压缩获得小型目标模型。在实践方面，在同样的参数量约束和训练配置下，减枝、量化、蒸馏得到的压缩模型究竟孰优孰劣是值得思考和探究的，目前还缺少不同压缩方法之间的全面而广泛的实验对比，这就导致在模型压缩方法的选择上会造成困惑。

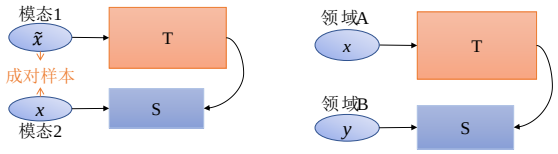
5.2 跨模态/跨领域

数据的存在形式称为模态(Modal)，在实际应用中，同一事物或事件的描述可能有多种形式，不同形式的数据可以丰富人们对事物的认知。因此，通过跨模态(Cross-modal) 的学习可以建立不同数据之间的关系，从而使得学习效果得到改进。

知识蒸馏可以很好的结合跨模态学习，其结构如下图 12(a)。在不同的模态学习中实现知识的提取、融合、转移，以及在各个领域的应用跨模态的知识蒸馏的研究已经逐渐兴起，例如通过跨模态实现视觉和听觉的融合[14-17]来辅助提升目标任务的学习效果。具体而言，SoundNet[14]中通过收集未标记的视频数据，利用时间和声音的自然同步来学习声学表征，通过教师-学生模型实现两者之间的知识迁移。文献[15,16]有着相近的思路，前者[15]中的教师模型通过提取视频中面部表情，而学生模型则学习视频中对应的声音。不同的是后者[16]中，让教师（ASR 模型）识别语音，而让学生（VSR 模型）来学习视频。知识蒸馏应用在跨模态行为识别领域也具有广泛的研究[167,168]。比如，Zhao 等人[167]提出的方法较为新颖，根据 WIFI 信号通过墙体并能在人体身上反射的事实，通过视觉和信号构建教师-

学生训练模型来评估人体的姿态，训练完成后无线信号模型能透过墙壁来估计人体的姿势。Thoker 等人[168]针对 3D 姿势的序列动作识别，通过提取源模态的教师网络的知识并将其转移到目标模态中，其中学生网络可以通过多个网络集成实现。类似地，Gupta 等人[169]在跨模态的图像识别任务上，将有监督的源图像和无监督的深度图和光流结合，通过原始的 RGB 图来学习丰富的语义表示并将学习到的“知识”用于指导无监督的深度图和光流图像的训练。值得关注的是，目前，视觉图像和自然语言结合的跨模态研究正越来越受到关注，其优势在于跨模态能够融合不同模态的特征，丰富对特征的表示，并且能产生意想不到的收益。比如说，文献[55]提出的跨模态散列(Cross-modal Hashing)方法，它能够将不同模式的内容，特别是视觉和语言上的内容映射到同一空间，利用教师-学生模型优化来传播知识，从而提高跨模态数据检索的效率。

跨模态可以实现视觉、文本、语音等不同领域的结合，对深度学习发展具有重要研究意义，但也同样面临着一些的困境。首先是多元化的数据获取存在着数据壁垒，跨模态的效果很大程度上依赖于多元化的数据表示；但是，一些关键领域的用户数据是不开放的或者是敏感的、隐私的，因而这部分数据很难获取。其次，对样本要求较为严格，需要成对样本之间具有无偏性，如常见的视频或语音存在掉帧或者噪声的情况会造成一定的缺失，这会对模型学习造成较大的影响。



(a) 跨模态知识蒸馏　　(b) 跨领域知识蒸馏

图 12 跨领域和跨模态模型结构对比。

跨领域(Cross-domain)和跨模态的不同点在于跨模态是利用同一数据的不同形态来表示学习样本以实现不同知识的交互融合，而前者则是在不同任务之上进行学习并上实现知识的迁移[18-21,54]。其原因在于神经网络在图像中学习到的低级特征（如边，角等信息）在不同的任务中是可以共享的。因此，通过跨领域可以进一步拓宽了知识表示的范围，增加了模型的灵活性，也能够为不同领域知识的融合提供了可能，其结构如图 12(b)所示。在实际中，神经网络通常需要大量的数据进行训练，但是在特定领域中的收集大量有标签的数据是非常困难的，现有的方法通常是通过在大量数据集上训练模型，然后在特定任务上进行微调（如 ImageNet）。而在微调的过程中源数据集通常会被丢弃，因此，Zhao 等人[18]提出了一种轻量化的软微调方法，即在源数据域微调而后将所学习到的“判别知识”，转移到目标域中来提高目标域的鲁棒性同时加速收敛。同样的在文献[19]中，通过注意力桥接网络将图像的前景先验数据从一个简单的单标签的源域转移到复杂的多标签数据的目标域中，从而显著改善注意力图，通过覆盖目标区域，能够有助于提高弱监督语义分割的性能，引导分类网络学习完整的视觉模式，从而提高泛能力。类似的做法还有文献[21]，通过教师模型在不重叠的类别样本上作预训练，而后将教师的 embedding 关系和模型的分类作为知识并蒸馏给在不同任务上训练的学生模型，从而促进学生模型的训练。

知识蒸馏结合跨领域能够很好地解决交叉任务和不同任务上知识的融合。通过重用跨任务模型的知识有助于提升目标域的泛化效果和鲁棒性。其存在的主要问题在于源域中的数据分布和目标域数据分布不一致，可能会带来一定的偏差，因此在迁移过程中需要考虑域适应的问题。特别地，最近笔者开展了一项利用知识迁移解决无源域适配问题的研究[170]，这是知识蒸馏在域适配场景中的新问题、新思路。

5.3 隐私保护

深度学习时代也同样面临着数据隐私安全方面的问题，在使用私人和敏感数据进行训练时，深度学习必须解决日益增长的隐私担忧。

传统的深度学习模型很容易受到隐私攻击。例如，攻击者可以从模型参数或目标模型中恢复个体的敏感信息。因此，出于隐私或机密性的考虑，大多数数据集都是私有的，不会公开共享。特别是在处理生物特征数据、患者的医疗数据等方面。而且，企业通常也不希望自己的私有数据被潜在竞争对手访问。因此，模型获取用于模型训练优质数据，并不现实。对于模型来说，既希望能访问这些隐私数据的原始训练集，而又不能将其直接暴露给应用。因而，可以通过教师-学生结构的知识蒸馏来隔离的数据集的访问。让教师模型学习隐私数据，并将知识传递给外界的模型。例如，Gao 等人[25]提出的知识转移结合了隐私保护策略，这个过程中教师模型访问私有的敏感数据并将学习到的知识传递

给学生，而学生模型不能公开获取数据但是可以利用教师模型的知识来训练一个可以公开发布的模型，以防止敏感的训练数据直接暴露给应用，同时保证模型的实用性。Cha 等人[26]针对传统强化学习中经验池 RM 组件所包含的所有的状态和操作策略可能会导致隐私泄漏的问题，提出具有高效通信和保护隐私的分布式强化学习框架-联邦强化蒸馏，可以有效保护应用数据的隐私和发布。

知识蒸馏应用于隐私保护的研究正在逐渐兴起，它能够结合目前存在的绝大多数模型以相对简单有效的方式解决了目前行业中的关键问题，不仅能够节约企业的成本还提高了整体的效率。因此，关于知识蒸馏和隐私保护的一些研究，如少样本、零样本等将会成为接下来的研究热点。

5.4 持续学习

持续学习(Continual Learning) 是指一个学习系统能够不断地从新样本中学习新的知识，并且保存大部分已经学习到的知识，其学习过程也十分类似于人类自身的学习模式。但是持续学习需要面对一个非常重要的挑战是灾难性遗忘，即需要平衡新知识与旧知识之间的关系。

知识蒸馏能够将已学习的知识传递给学习模型实现“知识”的增量学习(Incremental Learning)。比如，通过蒸馏的方法来解决目标检测的任务[22-24]。为了解决灾难性遗忘的问题，Shmelkov 等人[22]通过损失函数来平衡新类的预测和蒸馏损失，最小化来自原始网络和更新网络的旧类别之间的差异。同样的工作还有 Chen 等人[23]采用增量模型来学习特征并通过中间层的 Hint loss 来指导原始模型的训练，同时结合置信损失来提取初始模型的置信信息，可以有效避免在旧数据上所学“知识”的遗忘。Zhou 等人[24]提出的蒸馏策略引入了多模型蒸馏，直接指导模型从对应的教师中提取知识，并通过辅助蒸馏保持中间特征，同时结合剪枝来减小模型的内存占用，提高整体的效率。

目前的持续学习方法绝大部分都是处理图像分类和目标检测任务，而在 Michieli 等人[171]的工作中，将其正式引入语义分割；提出了四种蒸馏策略，分别在输出层、中间层、编码器和解码器上进行优化设计，并结合标准的交叉熵损失来优化新类的性能，同时保持旧类别的识别效果，该算法在语义分割上取得了一定的效果。另外，在文献[172]中利用 GANs 来解决生成模型的终身学习问题，提出了更为通用的框架来持续学习生成模型在不同条件下的图像生成，LifelongGAN 采用知识蒸馏的方法，将以前的网络中学习到的知识转移到新的网络中，这使得在终身学习环境中执行图像条件生成任务成为可能。

知识蒸馏在分类任务持续学习中取得了良好的效果,其挑战性在于目标检测同时存在目标分类和定位双重任务，直接的知识蒸馏方法还不能在目标检测任务的增量学习中提供满意的结果。因此，关于这方面任务还需要更多的研究。此外，如何减少不同增量步骤之间的混淆，尤其是在不访问之前数据的情况下，将是接下来需要探索的方向之一。

6 交叉领域

6.1 生成对抗网络

生成对抗网络(Generative Adversarial Networks, GANs) 最先由 Ian Goodfellow 提出[27]。该模型主要通过生成器和判别器相互对抗来实现数据的合成和判别，目前已广泛应用于超分辨[173,174]、风格迁移[175,176]和图像合成[177,178]等任务中并取得了很好的效果；但因其计算复杂、存储开销过大，很难直接应用在低功率设备（如移动设备）中。因此，对于 GANs 的模型压缩的研究非常有必要。而知识蒸馏和 GANs 结合较为简单且可以在 GAN 模型的不同模块上实现蒸馏，主要可以分为 3 类：对生成器蒸馏[28,179]，对判别器蒸馏[80,130]和同时对生成器和判别器蒸馏[180]，如图 13 所示。

在早期，知识蒸馏与 GANs 相结合主要是将“知识”提供给判别器来增强判别模型的判别效果并实现对生成模型的蒸馏[28,179,181]，如图 13(a)。Belagiannis 等人[28]提出的框架中，将教师模型和学生模型作为生成器，其中教师模型是预训练模型，学生模型和判别器是在训练过程优化产生的，该方法在训练过程中不需要标签且能够推广到不同的教师学生模型。类似的压缩方法的工作还有文献[179]，但是文献[28]的模型并没有真正的生成器，而献[179]中的教师生成器是通过预训练的 W-DCGAN 框架实现的。近些年，随着零样本学习研究的深入，研究人员通过 GANs 模型来辅助生成数据并将预训练的教师模型和学生模型作为共同判别器；学生模型在训练过程中更新参数，从而实现了对 GANs 中判别模型的压缩[80,130,180]。其中最具代表性的是 DAFL[130]模型；该模型将预训练好的教师模型作为固定的鉴别器，利用生成器将输入的噪声转化为图

像，并通过教师模型和学生模型共同判别生成器产

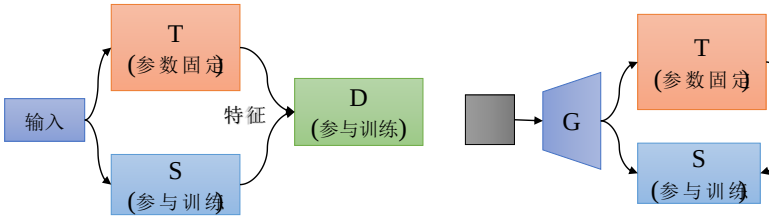


图 13 生成对抗网络结合知识蒸馏结构示意图（T 为教师模型，S 为学生模型，G 为生成器，D 为判别器）

生的图像；训练过程中教师的知识传递给学生模型，可以训练出一个轻量高效的学生模型，其结构如图 13(b)。类似的模型还有 DFAD[80]， DFAD 是在 DAFL 基础上做了近一步拓展，使得判别模型能适应不同的样本数据。与以上两种方法不同，Chen 等人[180]提出的模型较为特别，如图 13(c)。该模型同时在生成模型和判别模型上实现了蒸馏；其中生成模型由教师模型和学生模型共同组成用于图像的生成，而判别模型同样也是由教师模型和学生模型组成用于优化生成模型；该模型能够训练生成可移植性较强的生成模型。另外，在 GANs 的多模型结构中，知识蒸馏也具有很好的适应性。如 Chung 等人[182]提出的在线“知识”提取方法利用了对抗训练，不仅传递类别概率的知识，还传递特征图的知识；同时训练多个网络，将特征提取器作为生成器提取特征，使用判别器来区分不同网络中的特征图分布。同样地，Wang 等人[183]提出将知识从多个 GANs 转移到单个生成模型，利用教师网络和学生网络生成的特征映射分别作为真样本和假样本；并将两者进行对抗性训练来提高学生网络目标检测的性能。还有一些研究从不同的角度（鲁棒性[181]、判别边界[184]等）来提高模型泛化性，这也对 GANs 模型压缩提供了有力支撑。

目前，知识蒸馏结合 GANs 压缩上已经取得了很好的效果，但是在具体的应用中还存在着不易训练，不可解释等方面的挑战：

不易训练：主要表现在模型的稳定相对较差，根本原因在于 GANs 结构需要平衡多方且对于模型的参数较为敏感，训练过程中容易发生坍塌，尤其对一些复杂的任务（如细粒度、超分辨率等）具有较高的要求。

不易解释：现存的方法主要是经验性的结论缺乏一定的可解释理论，无法从理论上上解释在输出层蒸馏还是中间层蒸馏哪种更有优势。

蒸馏结合生成对抗网络虽然能在很大程度上实现模型的压缩，提升学生模型的性能，但是也无法从根本解决上述挑战。关于这方面的研究目前也是一个开放性的问题，期待更多的理论性研究被提出。

6.2 强化学习

强化学习(Reinforcement Learning, RL)，也被称为增强学习，主要用于描述和解决智能体在与环境的交互过程中通过学习策略以达到获取最大化回报的目标问题，如下图 14 所示。

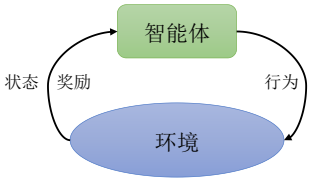
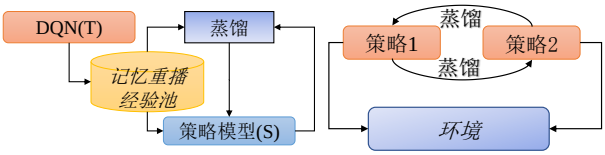


图 14 强化学习原理图。智能体在环境中根据观察的状态作为决策，采取相应的行为并期望获得最大的奖励。

目前，深度强化学习(Deep Reinforcement Learning, DRL) 模型的已经被广泛应用于各个领域[185]，如机器人控制[186]、完全信息博弈 [187]和非完全信息博弈 [188]。DRL 在应用过程中，深度模型需要通过与环境大量的交互获取奖励来更新智能体的网络参数，最终获得较高水平的表现。这使得模型的训练开销非常巨大。因此，研究人员将知识蒸馏应用于 DRL，期望辅助模型训练以提升模型的训练效果并且实现深度模型的轻量化。

知识蒸馏与深度强化相结合的过程中，研究人员最初提出了一种简单有效的方式被称为策略蒸馏[29,30,189]。该方法将预训练的深度 Q 网络(DQN)模型作为教师模型，并将其学习到的状态、经验和奖励作为知识存储到记忆重播(Replay Memory) 的经验池中，在训练学生模型时这些经验池中的“知识”将作为指导学生的依据，如图 15(a)。Rusu 等人[29]使用了监督回归训练学生模型，使其产生与教师模型相同的输出分布。该模型在一些具有挑战性的领域中取得了优异的表现；但是，由于预训练模型往往无法得到，因此获得教师策略的计算成本非常高；此外，如果教师模型不是最优的，学生模型的性能也会受到教师模型的限制。针对这一问题，

Lai 等人[190]构建了学生-学生框架的双策略蒸馏(Dual Policy Distillation, DPD)，如图 15(b)。在 DPD 框架中，两个学习模型在相同的环境中从不同的环境视角中探索并从彼此身上“蒸馏知识”来提高他们的学习效果。这样的模型能弥补教师模型的不足，平衡教师和学生模型之间的差异。其次，还可以采用集成策略的方式，如 Hong 等人[189] 提出的方法主要是周期性的将环境交互策略的集合作为知识，并通过 KD 在集合中的策略之间定期共享知识。



(a) 策略蒸馏 (b) 双策略蒸馏

图 15 强化学习中的知识蒸馏示意图。深度强化教师模型将经验值存到记忆重播池中，学生模型从策略池中学习教师模型的经验。双策略模型的两个模型从环境中学习经验并互相蒸馏知识。

策略蒸馏的主要挑战在于密集奖励和隐私安全。只有当环境中有密集的奖励时，模型通过随机行为才能更容易地获取奖励。如果环境中没有足够多的奖励，这种方法很容易失效。因此，探索奖励的蒸馏方法引起了研究人员的关注，如 Burda 等人[191]提出了将内部奖励与外部奖励相结合的随机网络蒸馏模型，使得模型训练易于实现并且增加了模型的灵活性。除此之外，策略蒸馏的局部记忆重播(Local Replay Memory) 模块存在安全隐私泄露的问题。该模块用来存储智能体的行为策略和被观察状态，如果不加保护，直接让智能体访问，容易造成隐私泄露等安全问题。针对这样的问题，Cha 等人[26]提出了通信效率高、保护隐私的分布式 RL 框架-联邦强化蒸馏。在联邦强化学习中每个智能体交换他们的经验重播记忆，这其中的策略值是局部策略的均值，而状态值是实际状态值的聚类。该方法避免了直接传递用户的行为和策略，起到了隐私保护的作用。此外，知识蒸馏结合深度强化学习还可以解决大型深度学习网络中复杂的搜索路径问题[91,152]。在一些分布式架构实现的强化学习网络中，也起到稳定和加速学习的效果[26,192]。

知识蒸馏应用于深度强化学习其重点在于能够保证模的压缩，较少的精度损失可以忽略。它能够较为灵活地解决其面临的困境，但其在应用中仍然有很多挑战性的问题值得探究。比如在一些复杂的任务中稀疏反馈问题严重阻碍了智能体的性能提升，本文认为该问题可以通过分层强化学习结合知识蒸馏实现提升智能体的层次化感知并在此基础上传递奖励机制促进探索。其次，目前知识蒸馏主要解决（单）智能体在单一任务和一些复杂场景下的多任务知识迁移，关于多智能体结合的复杂决策蒸馏[193,194]是接下来值得关注的研究方向之一。最后如何有效地将已知环境的探索和利用策略迁移到智能体与未知环境交互的应用上也是亟待解决的问题之一。

6.3 元学习

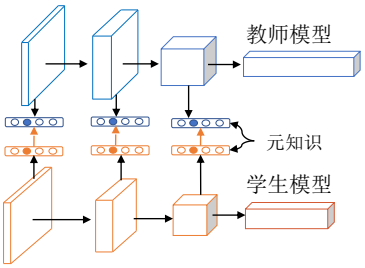


图 16 元学习知识蒸馏结构图。在教师和学生模型中构建“元知识”用于辅助学生训练。

元学习(Meta Leaning) 主要研究如何让机器学会学习。传统的机器学习和深度学习模型需要在大量样本上进行训练，这种方法缺点是在任何任务上都必须满足大量样本的前提。然而在很多现实领域，实际上无法获得到足够多的样本。而反观人类只需要在很少的样本上进行学习即可掌握知识。因而，很多学者认为元学习是实现通用人工智能的关键。

近些年，元学习逐渐成为研究的热点，主要应用在少样本(Few-shot) 分类任务上。它通过对少量的标签样本快速学习预测模型以适应新的类别，这其中最关键的是如何有效表示并传递“元知识”，其中“元知识”可以是参数、梯度、注意力等等。目前，很多研究[31,32]将其和知识蒸馏结合，通过在不同任务之间实现“元知识”传递和共享，以提高模型的泛化性能，其结构如图 16 所示。如 Dvornik 等人[32]通过额外的无标签数据以最小的代价损失将集成网络的知识蒸馏给单个网络使得整体性能得到提升。Li 等人[31]提出的少样本知识蒸馏能够从无标签的少数样本中提取知识，并将原始网络作为教师网络；教师网络通过剪枝或是分解后得到的压缩网络作为学生网络并通过最小二乘回归来实现教师和学生网络的输出拟合。该方法既能提高数据的有效性，又能提高训练/处理的效率。除了少样

本学习，元学习还可以结合表征实现知识的融合，使“知识”更加丰富，网络更加灵活、轻便[195-197]。

知识蒸馏结合的元学习作为小样本环境下提高性能的手段，在知识迁移过程中也会面临着一些挑战，诸如过拟合[33]、结构不匹配[196]、新旧任务不关联[198]等问题。

过拟合：主要是指压缩后的网络在少量的训练样本上容易产生过拟合，这会导致和原网络的估计误差较大，而且这种估计误差可能会逐层传播和累积，最终导致整个网络输出变差，针对这一挑战，传统的方法使用了中间层蒸馏来缩小教师-学生的差距。文献[33]通过层间交叉蒸馏作了相关的探索，虽然能进一步提高模型的泛化性能，但这种方法本质上是更进一步约束学生模型，虽然新颖但只适应于同构模型。本文认为合理数据增强、dropout、正则化等手段也能在一定程度上起到补充的作用，更进步的模型设计可以结合剪枝等算法。

结构不匹配：主要指异构网络之间层与层之间存在不匹配的问题。比如上文 3.2 提到通过适配器的方式实现层对层之间的传递知识，但实际情况是教师模型的学习能力较强，能够学习更抽象是高维特征而相同层的学生模型却有可能只学习到了低维特征，这样直接传递知识通常是不合理的；因此，在这种情况下，异构网络是很难将原网络的层和目标网络的层之间关联起来。文献[199]提出了基于元学习的迁移学习方法，该方法可以在给定原网络和目标网络的情况下，自动地挑选网络层中的重要特征。但是，如何设计算法来挑选匹配的特征仍然是值得探究的课题，相关研究也可以借鉴神经结构搜索来寻找最佳匹配的特征并建立合理评价机制。

新旧任务不匹配：通常情况下不同任务之间迁移需要满足任务之间有一定的相关性（具有一定相似性），前面在 5.2 节讨论不同任务之间的图像低维特征可以共享，因此特征迁移主要需要解决上述结构匹配的问题。但是在不同任务没有紧密关联的情况下传递原模型的参数知识的方法将面临困境，因此如何寻找到有效的参数是值得研究的课题。相关方法可以借鉴文献[198]给出的 Leap 模型，该模型将每个任务从初始化到最终参数的训练过程联系起来构成一个流形，并构建一个目标函数通过梯度求解使得这条路径长度最小化，在这个过程中将知识传递给学生模型。这种算法的优势在于训练过程中能够根据梯度来寻找出局部最优化的参数。

6.4 自动机器学习

自动机器学习(Auto Machine Learning, AutoML)是将机器学习整个流程通过端到端的方式实现自动化的过程。传统机器学习模型需要完成数据采集、数据预处理、模型优化和应用部署等步骤；而完成这些步骤需要花费大量的精力进行算法和模型的选择。AutoML 从传统机器学习模型出发，在特征工程、模型构建和超参优化三方面实现自动化并给出了端到端(End-to-End) 的解决方案[200]。在图像分类中，通常有两类 AutoML 技术，神经结构搜索(Neural Architecture Search, NAS)[34-36,196,201]和超参数优化[202] (Hyper-Parameter Optimization, HPO)，这两种方法都是利用自动化学习策略来代替人类经验。由于 NAS 在目前是研究的重点，因此，本文主要介绍 NAS 和知识蒸馏结合的研究。

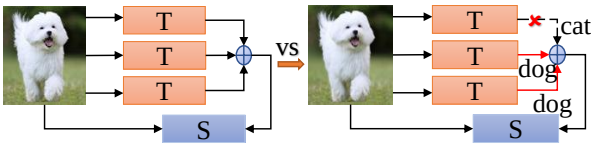


图 17 传统模型学习与自动机器学习对比图。左图为传统蒸馏从多个教师模型学习知识并集成，右图为自动机器学习，从教师模型学习并自动搜索出最准确的知识并集成。

NAS 是 AutoML 中用来自动识别深度神经网络中最优模型结构的技术，主要由搜索空间、搜索策略和性能评估策略三部分组成。搜索过程中，最精确的方法是在搜索空间中从零搜索，但是这种方法并不现实；其次，便是使用先进的搜索策略（如早期的进化算法或者目前流行的梯度优化和强化学习算法等），在子空间中搜索训练架构；这种方式成本仍然很高，需要很长的时间和资源的开销。因此，研究人员将知识蒸馏与 NAS 结合，在大型架构的网络空间中实现了与蒸馏结合的知识传递，如图 17 所示。如 Kang 等人[36]提出的智能体 Orcale 知识蒸馏，通过 NAS 来搜索并蒸馏有效的结构和操作，能够从集成教师中学习到强大而有效的学生模型。Macko 等人[203]提出的 AdaNAS，使用集成技术来自动地将 NAS 搜索到的神经网络组成一个更小的集成网络，并通过知识蒸馏将以前的集成模型作为教师对较小的网络进行迭代训练，在保持相同数量的参数的情况下，网络集成可以提高单个神经网络的精度。类似的工作还有文献[196,201]。

为了提高 NAS 的搜索速度，最近的研究[204,205]还提出了利用共享的网络参数在搜索空间中同时

训练不同的候选结构。然而，这导致了不正确的评估，从而降低了 NAS 的有效性。针对这一问题，Li 等人[206]提出将 NAS 的搜索空间模块化成块，以确保潜在的候选结构得到了充分的训练，并从教师模型中蒸馏神经结构知识作为监督来指导学生模型中每个块中的架构搜索，这很大程度上提高了 NAS 的有效性。值得一提的是，Dong 等人[207]为了突破剪枝网络的结构限制提出可变化结构搜索，该方法用 NAS 直接搜索具有最佳的通道和大小的网络作为学生网络，其中通道/层的数量是通过最小化剪枝网络的损失来学习的。

神经网络结构空间是非常巨大的，从庞大的神经元空间中搜索到最优的结构无疑是非常费时的。目前关于 NAS 方面的研究已经很多，对于 AutoML 的应用前景也更加的乐观，但是 AutoML 还存在一些争议，如文献[34,35]中认为很多存在的 NAS 解决方案并不一定比随机结构选择好。其次，NAS 结合知识蒸馏的过程中，还有一些需要解决的挑战的难题，包括结构不匹配、搜索空间复杂、鲁棒性不足等问题：

结构不匹配：教师模型和学生模型性能不匹配导致的学生模型无法有效学习的问题，这挑战类似 6.3 节中的结构不匹配问题，这里不再赘述；

搜索空间复杂：目前的应用中，针对不同的任务场景往往需要不同模型，也意味着具有不同的搜索空间。因此，如何高效的灵活的实现不同任务下的自动空间搜索和迁移从而摆脱人类经验也是一个值得探究的问题；

鲁棒性不足：AutoML 虽然在已有的一些数据集上取得很好的效果，但是这些数据集基本上是标签完备的，实际应用中可能会面临着较多的噪声，因此如何结合蒸馏有效避免噪声提高学习的鲁棒性也是非常有必要的；

此外，目前基于 AutoML 设计的蒸馏模型绝大部分是应用于计算机视觉领域，而其它的领域如自然语言处理中结合相对较少。因此，关于其他方向上的研究仍是值得期待的。

6.5 自监督学习

自监督学习(Self-Supervised Learning, SSL) 是介于监督学习和无监督学习之间的一种新的范式，旨在减少深度网络对大量标注数据的依赖。其主要思想是利用辅助任务从大规模的无监督数据中挖掘出数据本身的结构、语义、属性等信息，从而学习到对下游任务有价值的表征。

目前，自监督学习在预训练模型上的应用非常成功，其与监督学习预训练最大的不同在于：监督学习预训练主要是在大量有标签的数据上进行训

(a) 监督学习蒸馏　　(b) 自监督学习蒸馏

图 18 监督学习和自监督学习蒸馏结构对比图。传统的监督学习的蒸馏在标签数据集上构建预训练模型（标签任务），而自监督学习蒸馏则是在无标签数据集上训练并‘总结’出知识（辅助任务），用于目标模型的训练。

练得到预训练模型，而后将预训练模型应用于下游任务，通过微调(Fine-tuning) 的方式以适应新的任务，如图 18(a)；而自监督学习是在大量的无标签数据上通过构建辅助任务来训练模型，并将得到预训练模型通过微调的方式应用于下游任务，如图 18(b)。但是自监督学习这种预训练-微调方法的缺点在于学习辅助任务和目标任务时只能使用同构模型或者其中的一部分，这也导致了目前绝大部分自监督学习的方法在预训练和微调时都是使用的相同架构。因此，Noroozi 等人[37]为了克服这样的局限性，通过知识蒸馏方法解耦这两种体系结构，允许在预训练上使用更深的模型。

除了模型结构上的不足外，自监督学习的另外一个重要的挑战在于如何从大量的无标签数据中提取出丰富表征？为此，研究人员从不同的角度来探索其中的特征表示方法，这其中比较有代表性的工作有通过拼图的方式构建图像上下文信息(如 Jigsaw[68,69]) 或者是根据图像本身的信息(如角度[67]、颜色[70]等)，但是后来越来越多的工作开始思考自监督学习和具体任务紧密结合的方法。本文主要介绍自监督与知识蒸馏相结合的方法。为了从预训练教师模型中提取丰富信息同时解决教师-学生模型结构有限，Lee 等人[38]则通过奇异值分解(SVD)对数据进行压缩并利用径向基(RBF)网络来分析压缩特征之间的相关性；这种方法会自动化创建标签并持续使用，确保转移的知识不会消失。Lee 等人[67]提出将自监督学习和自蒸馏学习相结合，该方法不同于之前自监督任务，将数据增强学习的类别信息和多任务学习的角度信息融合成为类别-角度成对信息并通过自蒸馏将高维的监督信息传递给模型低维的空间用于增强模型表示能力。而 Xu 等人

[71]探索了一种更普遍的方法从预训练模型获取更丰富的知识。该方法在辅助任务中添加模块和分支将其更新为参数固定教师模型，并从教师模型输出的结果中构建一个正样本对和几个负样本，而后探索单个正样本和多个负样本之间的结构化知识；并将这种具有相似性的结构化知识传递给学生模型。这种方法在在小样本和噪声标签的场景下也具有很好的应用，类似的工作还有文献[72]。针对小样本学习，通过自监督知识蒸馏从中快速学习无序分布来提高小样本学习任务的模型表示能力。该方法主要分为两个阶段，在第一阶段通过构建自监督学习的辅助任务来最大化特征嵌入的熵，在第二阶段将辅助任务的中的参数分别克隆给教师和学生模型并在教师-学生模型之间的实现知识蒸馏。

以上提到的方法主要是研究样本内特征的表示，而样本间的特征其实也是具有很多约束关系的存在。比较有代表性的是时序约束，而这种约束关系也可以通过自监督学习的方法来表示。比如视频中利用自监督学习来学习相邻帧之间的相似性或者将同一物体的多个视角当作同一帧来学习特征的相似性[73,74]。但是这种方法很难扩展到目前具有海量数据的应用中，主要存在两个局限性：首先，目前的自监督学习方法多只关注于单个任务，而忽略了不同任务特征之间的互补性，从而导致视频表现不是最优的。其次，较高的计算和内存成本阻碍了它们在现实场景中的应用。Zhang 等人[105]提出的基于图的解决方法，通过从多个自监督任务来转移知识，这里的“知识”包括了输出层的 logits 图和中间层的特征图；前者蒸馏类别知识可以解决多元联合分布问题，后者蒸馏内部特征知识解决不同特征之间异质性的挑战。该方法可以显著减少从教师那里学到的冗余知识，使学生模型更轻更有效解决分类任务。

在信息化时代中，海量数据是相对容易获取的，但是人工标签的成本巨大，因此，这成为人工智能推广和发展的一大阻碍。目前，自监督学习主要应用在标准数据集上取得了很好的效果。而现实中数据集的来源是非常多样的，因此，需要克服的挑战也是多样的。比如，从合成数据中学习知识、从网络数据（视频、图片、网页）中学习、从多个辅助任务中学习知识、还有一些特殊场景如传感器网络中学习等等。在这些复杂条件下如何有效学习知识并结合知识蒸馏是具有重要意义且值得探究的课题。

7 主要应用

7.1 计算机视觉

计算机视觉一直是人工智能的研究热点领域之一。近年来，知识蒸馏被广泛应用于各种视觉任务达到模型压缩、迁移学习和隐私保护等目标。虽然知识蒸馏的应用十分广泛，但是由于各个研究方向的热度不同，所以相关研究的论文数量也会有很大的差异。本文重点引用了知识蒸馏在视觉上的热点方向，并列举相关论文的方法供读者查阅学习，而对于其它一些方向可能存在取舍。目前，应用知识蒸馏的视觉研究主要集中在视觉检测和视觉分类上。视觉检测主要有目标检测、人脸识别、行人检测、姿势检测；而视觉分类的研究热点主要是语义分割，如表 5 所示。另外，视觉中还有一些其它应用比如视频分类[105]、深度估计和光流/场景流估计[169]等等。

7.2 自然语言处理

自然语言处理(Natural Language Process, NLP)的发展非常迅速，从 RNN，LSTM，ELMo 再到如今非常热门的 BERT，其模型结构逐渐变的非常的深而复杂，需要耗费的大量的资源和时间。这样的模型几乎无法直接部署。因而，获得轻量级、高效、有效的语言模型显得极为迫切。于是，知识蒸馏在 NLP 领域也得到了极大的重视。目前，结合知识蒸馏较为广泛的 NLP 任务主要有机器翻译(Neural Machine Translation, NMT)，问答系统(Question Answer System, QAS) 等领域。如表 6，本节列举了知识蒸馏结合神经机器翻译和问答系统的代表性的研究工作。另外，BERT 模型在近些年被广泛应用 NLP 的各个领域，其重要性不言而喻，因此，我们在表 6 中一并列举并在下面对其作详细介绍。

BERT 模型是近年来自然语言中，应用最广泛的工具之一，它是由双向编码器表示的 Transformer 模型组成。由于其强大的编码表示能力，目前在自然语言的各个任务中被广泛应用。但是，BERT 模型结构非常复杂，参数量巨大，很难直接应用于模型的训练。目前的应用主要采用的预训练加微调的方法，因此，对 BERT 模型的压缩显得尤为必要。目前，这方面的研究已经吸引的很多研究者的关注。提出的方法主要有剪枝、量化、蒸馏、参数共享、权重分解。但是，量化对模型的提升效果有限，权重分解和参数共享等工作相对较少。因此，主要

表 5 计算机视觉主要蒸馏方法应用与对比（注：‘A’表示离线蒸馏，‘B’表示在线蒸馏，‘C’表示自蒸馏，‘D’表示无数据蒸馏，‘E’表示多模型蒸馏，‘F’表示特权蒸馏；‘L’表示标签知识，‘I’表示中间层知识，‘P’表示参数知识，‘S’表示结构知识；‘M’表示模型压缩，‘K’表示跨模态/领域，‘H’表示隐私保护，‘J’表示持续学习，下同。）

任务	作者	时间	蒸馏范式						知识形式				蒸馏目的			
			A	B	C	D	E	F	L	I	P	S	M	K	H	J
目标检测	Shmelkov 等人[22]	2017	✓						✓							✓
	Wei 等人[2]	2018	✓							✓			✓			
	Chen 等人[1]	2019		✓					✓	✓			✓			
	Wang 等人[3]	2019		✓						✓						✓
	Tang 等人[83]	2019		✓						✓			✓			
人脸识别	Ge 等人[208]	2018	✓					✓	✓	✓	✓	✓	✓			
	Kong 等人[209]	2019	✓							✓				✓		
	Yan 等人[210]	2019	✓		✓		✓			✓		✓	✓			
	Karlekar 等人[211]	2019	✓							✓	✓		✓			
	Wu 等人[212]	2020	✓						✓	✓		✓	✓			
行人检测	Shen 等人[213]	2016	✓						✓	✓	✓		✓			
	Kruthiventi 等人[214]	2017	✓							✓				✓		
	Chen 等人[215]	2019		✓					✓	✓			✓			
姿势检测	Angel 等人[216]	2019	✓						✓	✓			✓			
	Nie 等人[217]	2019		✓	✓					✓	✓		✓			
	Wang 等人[218]	2019	✓							✓	✓		✓	✓		
	Hwang 等人[219]	2020	✓						✓	✓			✓			
	Zhang 等人[220]	2020		✓	✓					✓			✓	✓		

语义分割	Chen 等人[81]	2018	✓							✓				✓	✓			
	Fang 等人[80]	2019	✓			✓			✓					✓			✓	
	He 等人[79]	2019	✓						✓					✓	✓			
	Liu 等人[78]	2019	✓						✓	✓		✓		✓	✓			
	Dou 等人[82]	2020		✓			✓		✓	✓				✓	✓			

工作集中在剪枝和蒸馏。此处将主要介绍表中列举的较为经典的几种模型。首先，知识蒸馏结合 BERT 较早的方法是 Distilled BiLSTM[221]于 2019 年提出，其主要思想是将 BERT-large 蒸馏到了单层的 BiLSTM 中，其效果接近 EMLO，其将速度提升 15 倍的同时使模型的参数量减少 100 倍。后来的研究方法逐渐丰富，如 BERT-PKD[222]主要从教师的中间层提取丰富的知识，避免在蒸馏最后一层拟合过快的现象。DistillBERT[223]在预训练阶段进行蒸馏，能够将模型尺寸减小了 40% ，同时能将速度能提升 60%，并且保留教师模型 97%的语言理解能力，其效果好于 BERT-PKD。TinyBERT[224]提出的框架，分别在预训练和微调阶段蒸馏教师模型，得到了速度提升 9.4 倍但参数量减少 7.5 倍的 4 层 BERT，其效果可以达到教师模型的 96.8%。同样，用这种方法训出的 6 层模型的性能超过了 BERT-PKD 和 DistillBERT，甚至接近 BERT-base 的性能。上述介绍的几种模型都利用了层次剪枝结合蒸馏的操作。MobileBERT[225]则主要通过削减每层的维度，在保留 24 层的情况下，可以减少了 4.3 倍的参数的同时提升 4 倍速度。在 GLUE 上也只比 BERT-base 低了 0.6 个点，效果好于 TinyBERT 和 DistillBERT。此外，MobileBERT 与 TinyBERT 还有一点不同，就是在预训练阶段蒸馏之后，直接在有一点不同，就是在预训练阶段蒸馏之后，直接在

表 6 自然语言处理的主要蒸馏方法应用与对比

任务	作者	时间	蒸馏范式						知识形式				蒸馏目的			
			A	B	C	D	E	F	L	I	P	S	M	K	H	J
机器翻译	Kim 等人[4]	2016	✓						✓	✓			✓	✓		
	Freitag 等人[5]	2017	✓							✓			✓			
	Hahn 等人[6]	2019			✓				✓			✓	✓			
	Zhou 等人[226]	2019	✓				✓			✓			✓			
	Tan 等人[227]	2019	✓				✓		✓				✓			
	Wei 等人[228]	2019		✓			✓		✓	✓			✓			
	Gordon 等人[229]	2019	✓						✓	✓			✓	✓		

问答系统	Wang 等人[230]	2018	✓					✓	✓				✓		
	Hu 等人[231]	2018	✓				✓	✓	✓	✓		✓		✓	
	Arora 等人[232]	2019	✓					✓				✓			
	Yang 等人[233]	2020	✓				✓	✓				✓			
BERT 模型	Tang 等人[221]	2019	✓					✓				✓			
	Sun 等人[222]	2019	✓					✓	✓			✓			
	Sanh 等人[223]	2020	✓					✓				✓			
	Mukherjee 等人[234]	2020	✓					✓	✓			✓	✓		
	Zhao 等人[235]	2020		✓				✓	✓	✓		✓			
	Jiao 等人[224]	2020	✓					✓	✓	✓		✓			
	Sun 等人[225]	2020	✓					✓	✓	✓		✓			
	Xu 等人[236]	2020		✓	✓			✓	✓			✓			
	Liu 等人[237]	2020			✓			✓	✓			✓			

MobileBERT 上用任务数据微调，而不需要再进行微调阶段的蒸馏，更加便捷。

综上，BERT 压缩在近些年的发展还是较为显著的。这些方法对后 BERT 时代出现的大型预训练模型的如 GPT 系列等单向或双向 Transformer 模型的压缩具有很大借鉴意义。

7.3 推荐系统

近些年，推荐系统(Recommender Systems, RS)被广泛应用于电商、短视频、音乐等系统中，对各个行业的发展起到了很大的促进作用。推荐系统通过分析用户的行为，从而得出用户的偏好，为用户推荐个性化的服务。因此，推荐系统在相关行业中具有很高的商业价值。深度学习应用于推荐系统同样面临着模型复杂度和效率的问题。但是，目前关于推荐系统和知识蒸馏的工作还相对较少。本文在表 7 中整理了目前收集到的相关文献，可供研究人员参考。

7.4 其它

对于人工智能的其它领域的应用。为了提高模型的训练的效率，知识蒸馏逐渐被尝试应用在其模型的设计中以获取轻量级深度声学模型，比如语音增强[243]、语音合成[244]、语音识别[245-247]等等。另外，由于开源社区的驱动与贡献，诞生了许多成熟的蒸馏与压缩的代码工具箱，整合了多种蒸馏算法，以帮助研究人员更快地进行迭代开发与研究工作。Intel 实验室推出的 Distiller[248]是一个开源的、支持知识蒸馏、剪枝、量化和稀疏分解等模型压缩范式的框架，并且支持图像分类、目标检测、语言处理、机器翻译、推荐系统等多种任务。该框架还提供了基于 PyTorch[249]的完整说明文档和使用教

程。哈工大与科大讯飞联合开发了专门针对自然语言处理的知识蒸馏工具箱 Textbrewer[250]。torchdistill[251]是一款基于配置文件的知识蒸馏框架，主要支持常用的图像分类、检测、分割模型。

表 7 推荐系统中的主要蒸馏方法应用与对比

任务	作者	时间	蒸馏范式						知识形式				蒸馏目的			
			A	B	C	D	E	F	L	I	P	S	M	K	H	J
推荐系统	Zhou 等人[238]	2018		✓					✓	✓			✓			
	Chen 等人[7]	2018	✓							✓			✓		✓	
	Tang 等人[8]	2018	✓						✓				✓			
	Pan 等人[9]	2019	✓						✓				✓			
	Xu 等人[239]	2020	✓	✓					✓				✓		✓	
	Mi 等人[240]	2020		✓					✓							✓
	Zhu 等人[241]	2020	✓	✓			✓		✓	✓			✓			
	Kang 等人[242]	2020	✓						✓	✓			✓			

KD-Lib[252]与 Distiller 类似，提供了较为完善的说明文档，支持蒸馏、剪枝与量化。还有一些个人开发者贡献的蒸馏工具箱，如 Knowledge-Distillation-Zoo、RepDistiller、classification distiller，都是整合与统一了多种蒸馏算法。这些代码工具箱或框架的发布与开源，极大地推动了知识蒸馏方面的研究进展，并且提供了统一的参考标准。

8 原理和解释

在联结主义人工智能时代，深度学习模型可解释性[41,42,253-255]越来越受到质疑和重视。关于知识蒸馏的原理也存在非常大的争论，根本上还是因为深度神经网络模型的可解释性存在困难。但是，目前已经有很多学者尝试从软标签正则化与泛化性角度探寻知识蒸馏的原理和可解释性，并且取得了一些成果。这对于增强知识蒸馏算法的可靠性，研究新的更高效的蒸馏方式奠定了理论基础，并且极大地推动了相关应用的发展。

关于软标签在知识蒸馏中作用机理的研究，常常与经典的标签平滑正则化(Label Smoothing Regularization, LSR)[41,62,256,257]联系在一起。Müller 等人[41]通过一种新的可视化方法研究了网络倒数第二层的表征，实验指出标签平滑会促使网络倒数第二层的激活值更接近真实类别的类中心并且与其他非正确类别的类中心的距离是相等的。并且，其证明了使用准确率高的教师模型对于获得更好蒸馏效果来说并不是必须的。如果教师是用标签平滑训练的，那么它指导的学生模型的效果反而会变差。标签平滑的作用是将相同类别的训练样本聚类到一个紧凑的集合中，但是这也导致了不同类别样本之间的相似性信息的缺失，从而影响了模型蒸馏的效果。Cho 和 Hariharan [258]的工作关注 KD 的有效性，得到的结论是：教师网络精度越高，并不意味着学生网络精度越高。这个结论和 MirzadehS-I 等人[48]的工作是一致的。较大的性能和容量差距使得学生网络不能稳定地模仿教师网络。有趣的是，Cho 和 Hariharan [258]认为多步蒸馏并非有效，提出应该采取措施提前停止教师网络的训练。

Liang 等人[257]首次提出了两个 DNN 中间层

之间知识一致性的定义，开发了一种与任务无关的方法，以从中间层特征中解开并量化不同顺序的一致特征。Cheng 等人[259]研究的核心在于通过定义并量化神经网络中层特征的“知识量”。从神经网络表达能力的角度来解释知识蒸馏算法的成功机理，提出并验证了三个假设：知识蒸馏使 DNN 比从原始数据中学习更多的视觉概念；知识蒸馏保证了 DNN 易于同时学习各种视觉概念；知识蒸馏比从原始数据中学习得到更稳定的优化方向。Rahbar 等人[260]证明了有关学生网络学习内容以及收敛速度的结果。Menon 等人[261]提出了关于蒸馏的统计学观点，并拓展了一种在多类检索中的新应用。Jha 等人[262]研究了 KD 方法在训练没有任何残差连接的深层网络中的功效。发现在大多数情况下，非残差学生网络的性能与不使用 KD 训练的残差版本的性能相同或更好。在某些情况下，即使是利用性能较差的教师训练学生也会提升性能，这与文献[62]的研究是一致的。文献[263]从理论上分析了神经网络的知识蒸馏：首先，为网络的线性化模型提供了转移风险界限；然后，提出一项衡量任务训练难度的数据无效性指标，并根据此指标表明，对于性能良好的教师较高比例的软标签对蒸馏是有益的；最后，对于教师不完善的情况，发现硬标签可以纠正教师的错误预测。这说明了混合硬标签和软标签的有效性。文献[264]中介绍了一个统计物理框架。该框架可以对浅层神经网络中的知识蒸馏属性进行表征分析。另外，较大的教师模型的正则化属性可以通过蒸馏由较小的学生模型继承，并且所产生的泛化性能与教师的最优性密切相关并受其限制。

另外，关于知识蒸馏的泛化理论也是一直以来的研究热点。传统研究中关注的重点在于结构复杂性和样本复杂性对于泛化能力的影响。知识蒸馏中的泛化原理主要研究如何通过蒸馏的泛化边界来提高网络的泛化性能[42,265-268]。如 Gong 等人[265]以特权信息为例分析了泛化蒸馏的原理在于软标签为学生提供了本身不包含的类别先验，使得学习到的分布偏向于教师的输出，同时有效降低了学生的假设空间。而后他们从理论上推导了半监督学习泛化蒸馏在无标签训练样本上的传导误差边界和测试样本上的归纳误差边界都是有上界的。文献[42]中，Phuong 和 Lampert 证明了泛化边界建立起了蒸馏训练线性分类器的期望风险的快速收敛条件，其关键因素在于数据分布的几何属性（特别是类别距离）、优化偏置（保证梯度下降能有最优路径）以及强单调性（增加数据集以减少学生分类器风险）。Hsu 等人[266]通过泛化理论分析表明假设使用较好的数据增强，原网络可以通过蒸馏继承好的泛化性。此外，蒸馏不仅能够提升泛化上界，而且能够有效的捕捉预测模型的内在属性，换言之可以在蒸馏过程中有效避免影响泛化上界的一些不好的依赖。类似地，Zhu 等人[268]在无数据蒸馏的异构联邦学习中通过理论推导证明了使用与全局分布一致的增强数据，可以提高模型的泛化性能。值得一提的是，Mobahi 等人[265]从理论上分析了自蒸馏的隐式正则化效果，即自蒸馏通过逐步自我迭代可以限制在基函数解的数量来改进正则化；并且在经验上证明了少量的自蒸馏可能会减少过拟合，进一步的自蒸馏可能会导致拟合不足，从而导致性能下降。虽然其工作是针对隐式正则化效果的解释，但是也能转化为对泛化边界的解释，这也就意味着泛化边界和正则化存在某种内在的联系。

关于知识蒸馏原理和解释的工作正在不断完善和丰富。相信这些研究会大大促进基础方法和下游应用的创新。通过以上阐述可以发现，目前关于知识蒸馏的解释主要集中在软标签正则化以及泛化边界的角度，而对于中间层、结构化等知识的解释十分有限，对于多种蒸馏学习方式的原理探究也有待关注。

9 未来发展趋势

根据对知识蒸馏领域的文献整理发现，知识蒸馏自 2015 年[10]被提出以来，在蒸馏方法上历经探索和改进，随后在各种应用领域开始拓展，在 2019-2020 年相关工作出现井喷式发展，文章数量占据了 80% 以上。相对地，蒸馏方法的改进工作趋于完善，多种任务、多种场景下的相关应用大量涌现，这与人工智能的发展趋势是分不开的。一是移动计算、物联网等人工智能应用落地场景下对深度神经网络模型体量和计算量的严格限制，对模型压缩产生了大量需求，而知识蒸馏具备模型压缩的天然特性。二是图神经网络[107-112]、BERT[221,237]、生成对抗网络[128,179,180,182,183,269]、胶囊网络[270,271]、视觉 Transformer[272-274]等一系列新型网络的诞生和发展，对特定领域的问题有了越来越完备的解决方案。三是神经网络可解释性和原理的研究日趋受到关注，因为这决定了目前的人工智能技术应用的可靠性和安全性。四是自监督学习[37,38,68-75]、因果推

断[275,276]等新的学习视角代表了人类对迈向真正人工智能的道路探索。基于此，本文对知识蒸馏这类新兴的学习方法的未来发展趋势作以下分析：

1) **模型压缩方面：**知识蒸馏对于压缩模型的方式主要是提升学生模型的性能、尽可能缩小与教师模型之间的差距，而不能真正地压缩目标模型的体量。剪枝、量化等模型压缩范式都可以与知识蒸馏无缝衔接，在直接缩减模型体量的同时，避免性能的衰减。目前缺少针对剪枝和量化范式的定制化蒸馏方法，因为剪枝要考虑模型结构的动态变化，量化需要考虑数值表示和表征空间的差异。
2) **新领域新场景：**随着新型网络结构的诞生与应用，势必都要面对性能与资源约束这一矛盾问题，也会需要考虑隐私保护和迁移学习等，这些都为知识蒸馏开拓了新的应用场景。在一些高层次任务上，如人体姿态估计、异常监测等，知识蒸馏的运用还只是初步阶段，还需要根据特定的领域知识和任务特点设计有针对性的蒸馏方法。
3) **原理与可解释性：**相关学者对于神经网络的原理和可解释性的追求从未间断，而对于知识蒸馏的原理的相关研究也将推动更高效的蒸馏方法的产生。结合传统机器学习中一些具备良好可解释性的建模方式，比如从模型不确定性、能量与信息论等角度，可以开展对蒸馏学习理论的研究。
4) **新的探索：**自监督学习、因果推断为当下的深度神经网络模型注入了新的活力，知识蒸馏也需要从新的视角探索开拓用武之地。自监督学习的内在关系表征在神经网络模型之间的迁移是十分值得探索的；利用因果推断去混杂影响也会改变知识蒸馏的思路。

10 总结

近年来，知识蒸馏逐渐成为研究热点而目前绝大多数优秀的论文都是以英文形式存在，关于系统性介绍知识蒸馏的中文文献相对缺失；并且知识蒸馏发展过程中融入了多个人工智能领域，相关文献纷繁复杂，不易于研究人员对该领域的快速、全面地了解。鉴于此，本文对知识蒸馏的相关文献进行了分类整理和对比，并以中文形式对知识蒸馏领域的研究进展进行了广泛而全面的介绍。首先介绍了知识蒸馏的背景和整体框架。然后分别按照知识传递的形式、学习方式、学习目的、交叉领域的结合对知识蒸馏的相关工作进行了分类介绍和对比，分析了各类方法的优缺点和面临的挑战，并对研究趋势提出了见解。本文还从计算机视觉、自然语言处理和推荐系统等方面概述了知识蒸馏在不同任务和场景的具体应用，对知识蒸馏原理和可解释性的研究进行了探讨。最后，从 4 个主要方面阐述了对知识蒸馏未来发展趋势的分析。

知识蒸馏通过教师-学生的结构为深度神经网络提供了一种新的学习范式，实现了信息在异构或同构的不同模型之间的传递。不仅能够帮助压缩模型和提升性能，还可以联结跨域、跨模态的知识，同时避免隐私数据的直接访问，在深度学习背景下的多种人工智能研究领域具有广泛的应用价值和研究意义。目前，有关知识蒸馏的中文综述性文章还比较缺失。希望本文对知识蒸馏未来的研究提供有力的借鉴和参考。

参 考 文 献

[1] Chen G, Choi W, Yu X, et al. Learning efficient object detection models with knowledge distillation//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 742-751.

[2] Wei Y, Pan X, Qin H, et al. Quantization mimic: Towards very tiny cnn for object detection//Proceedings of the 15th European conference on computer vision. Munich, Germany, 2018: 267-283.

[3] Wang T, Yuan L, Zhang X, et al. Distilling object detectors with fine-grained feature imitation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4933-4942.

[4] Kim Y, Rush A M. Sequence-level knowledge distillation//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Texas, USA, 2016: 1317-1327.

[5] Freitag M, Al-Onaizan Y, Sankaran B. Ensemble distillation for neural machine translation. arXiv preprint arXiv:1702.01802, 2017.

[6] Hahn S, Choi H. Self-knowledge distillation in natural language processing. arXiv preprint arXiv:1908.01851, 2019.

[7] Chen X, Zhang Y, Xu H, et al. Adversarial distillation for efficient recommendation with external knowledge. ACM Transactions on Information Systems, 2018, 37(1):1-28.

[8] Tang J, Wang K. Ranking distillation: Learning compact ranking models with high performance for recommender system//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London, UK, 2018: 2289-2298.

[9] Pan Y, He F, Yu H. A novel enhanced collaborative autoencoder with knowledge distillation for top-n recommender systems. Neurocomputing, 2019, 332:137-148.

[10] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

[11] Romero A, Ballas N, Kahou S E, et al. Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550, 2014.

[12] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928, 2016.

[13] Kim J, Park S, Kwak N. Paraphrasing complex network: network compression via factor transfer//Proceedings of the 32th International Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 2765-2774.

[14] Aytar Y, Vondrick C, Torralba A. Soundnet: Learning sound representations from unlabeled video. Advances in neural information processing systems, 2016, 29:892-900.

[15] Albanie S, Nagrani A, Vedaldi A, et al. Emotion recognition in speech using cross-modal transfer in the wild//Proceedings of the 26th ACM international conference on Multimedia, Seoul, Korea, 2018: 292-301.

[16] Afouras T, Chung J S, Zisserman A. Asr is all you need: Cross-modal distillation for lip reading//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain, 2020: 2143-2147.

[17] Fayek H M, Kumar A. Large scale audiovisual learning of sounds with weakly labeled data. arXiv preprint arXiv:2006.01595, 2020.

[18] Zhao Z, Zhang B, Jiang Y, et al. Effective domain knowledge transfer with soft fine-tuning. arXiv preprint arXiv:1909.02236, 2019.

[19] Li K, Zhang Y, Li K, et al. Attention bridging network for knowledge transfer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 5198-5207.

[20] Zhao S, Wang G, Zhang S, et al. Multi-source distilling domain adaptation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 12975-12983.

[21] Ye H J, Lu S, Zhan D C. Distilling cross-task knowledge via relationship matching//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,USA, 2020: 12396-12405.

[22] Shmelkov K, Schmid C, Alahari K. Incremental learning of object detectors without catastrophic forgetting//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 3400-3409.

[23] Chen L, Yu C, Chen L. A new knowledge distillation for incremental object detection//2019 International Joint Conference on Neural Networks. Budapest, Hungary, 2019: 1-7.

[24] Zhou P, Mai L, Zhang J, et al. M2kd: Multi-model and multilevel knowledge distillation for incremental learning. arXiv preprint arXiv:1904.01769, 2019.

[25] Gao D, Zhuo C. Private knowledge transfer via model distillation with generative adversarial networks. arXiv preprint arXiv:2004.04631, 2020.

[26] Cha H, Park J, Kim H, et al. Proxy experience replay: Federated distillation for distributed reinforcement learning. IEEE Intelligent Systems, 2020, 35(4):94-101.

[27] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montréal, Canada, 2014: 2672-2680.

[28] Belagiannis V, Farshad A, Galasso F. Adversarial network compression//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 431-449.

[29] Rusu A A, Colmenarejo S G, Gulcehre C, et al. Policy distillation. arXiv preprint arXiv:1511.06295, 2015.

[30] Czarnecki W M, Pascanu R, Osindero S, et al. Distilling policy distillation//The 22nd International Conference on Artificial Intelligence and Statistics. Macao, China, 2019: 1331-1340.

[31] Li T, Li J, Liu Z, et al. Few sample knowledge distillation for efficient network compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,USA, 2020: 14639-14647.

[32] Dvornik N, Schmid C, Mairal J. Diversity with cooperation: Ensemble methods for few-shot classification//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 3723-3731.

[33] Bai H, Wu J, King I, et al. Few shot network compression via cross distillation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 3203-3210.

[34] Yu K, Sciuto C, Jaggi M, et al. Evaluating the search phase of neural architecture search. arXiv preprint arXiv:1902.08142, 2019.

[35] Yang A, Esperança P M, Carlucci F M. Nas evaluation is frustratingly hard. arXiv preprint arXiv:1912.12522, 2019.

[36] Kang M, Mun J, Han B. Towards oracle knowledge distillation with neural architecture search//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 4404-4411.

[37] Noroozi M, Vinjimoor A, Favaro P, et al. Boosting self-supervised learning via knowledge transfer//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9359-9367.

[38] Lee S H, Kim D H, Song B C. Self-supervised knowledge distillation using singular value decomposition//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 335-350.

[39] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey. International Journal of Computer Vision, 2021, 129(6):1789-1819.

[40] Wang L, Yoon K J. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2021 (01): 1-1.

[41] Müller R, Kornblith S, Hinton G. When does label smoothing help? arXiv preprint arXiv:1906.02629, 2019.

[42] Phuong M, Lampert C. Towards understanding knowledge distillation//Proceedings of the 36th International Conference on Machine Learning, Long Beach, USA, 2019: 5142-5151.

[43] Li Y, Huang D, Qin D, et al. Improving object detection with selective self-supervised self-training//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 589-607.

[44] Xie Q, Luong M T, Hovy E, et al. Self-training with noisy student improves imagenet classification//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10687-10698.

[45] Ge Y, Chen D, Li H. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. arXiv preprint arXiv:2001.01526, 2020.

[46] Pham H, Dai Z, Xie Q, et al. Meta pseudo labels//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 11557-11568.

[47] Gao M, Shen Y, Li Q, et al. An embarrassingly simple approach for knowledge distillation. arXiv preprint arXiv:1812.01819, 2018.

[48] Mirzadeh S I, Farajtabar M, Li A, et al. Improved knowledge distillation via teacher assistant//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York City, USA, 2020: 5191-5198.

[49] Yang C, Xie L, Su C, et al. Snapshot distillation: Teacher-student optimization in one generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2859-2868.

[50] Müller R, Kornblith S, Hinton G. Subclass distillation. arXiv preprint arXiv:2002.03936, 2020.

[51] Guo J, Chen M, Hu Y, et al. Spherical knowledge distillation. arXiv preprint, 2020: arXiv: 2010.07485.

[52] Zhang Y, Lan Z, Dai Y, et al. Prime-aware adaptive distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 658-674.

[53] Wu G, Gong S. Peer collaborative learning for online knowledge distillation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event, 2021.

[54] Kundu J N, Lakkakula N, Babu R V. Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 1436-1445.

[55] Hu H, Xie L, Hong R, et al. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3123-3132.

[56] You S, Xu C, Xu C, et al. Learning from multiple teacher networks//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 1285-1294.

[57] Liu Y, Zhang W, Wang J. Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing, 2020, 415:106-113.

[58] Son W, Na J, Choi J, et al. Densely guided knowledge distillation using multiple teacher assistants//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual Event, 2021: 9395-9404.

[59] Shi W, Ren G, Chen Y, et al. Proxylesskd: Direct knowledge distillation with inherited classifier for face recognition. arXiv preprint arXiv:2011.00265, 2020.

[60] Wu A, Zheng W S, GUO X, et al. Distilled person re-identification: Towards a more scalable system//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1187-1196.

[61] Shin M. Semi-supervised learning with a teacher-student network for generalized attribute prediction//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 509-525.

[62] Yuan L, Tay F E, Li G, et al. Revisiting knowledge distillation via label smoothing regularization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 3903-3911.

[63] Zhang L, Song J, Gao A, et al. Be your own teacher: Improve the performance of convolutional neural networks via self distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 3713-3722.

[64] Yun S, Park J, Lee K, et al. Regularizing class-wise predictions via self-knowledge distillation//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, USA, 2020: 13876-13885.

[65] Phuong M, Lampert C H. Distillation-based training for multi-exit architectures//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 1355-1364.

[66] Hou Y, Ma Z, Liu C, et al. Learning lightweight lane detection cnns by self attention distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 1013-1021.

[67] Lee H, Hwang S J, Shin J. Self-supervised label augmentation via input transformations//Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020: 5714-5724.

[68] Doersch C, Gupta A, Efros A A. Unsupervised visual representation learning by context prediction//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1422-1430.

[69] Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles//Proceedings of the 16th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 69-84.

[70] Zhang R, Isola P, Efros A A. Split-brain autoencoders: Unsupervised learning by cross-channel prediction//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1058-1067.

[71] Xu G, Liu Z, Li X, et al. Knowledge distillation meets self-supervision//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 588-604.

[72] Rajasegaran J, Khan S, Hayat M, et al. Self-supervised knowledge distillation for few-shot learning. arXiv preprint arXiv:2006.09785, 2020.

[73] Wang X, Gupta A. Unsupervised learning of visual representations using videos//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2794-2802.

[74] Sermanet P, Lynch C, Chebotar Y, et al. Time-contrastive networks: Self-supervised learning from video//2018 IEEE International Conference on Robotics and Automation. Brisbane, Australia, 2018: 1134-1141.

[75] Si C, Nie X, Wang W, et al. Adversarial self-supervised learning for semi-supervised 3d action recognition//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 35-51.

[76] Ding Q, Wu S, Sun H, et al. Adaptive regularization of labels. arXiv preprint arXiv:1908.05474, 2019.

[77] Liu Y, Zhang W, Wang J. Learning from a lightweight teacher for efficient knowledge distillation. arXiv preprint arXiv:2005.09163, 2020.

[78] Liu Y, Chen K, Liu C, et al. Structured knowledge distillation for semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2604-2613.

[79] He T, Shen C, Tian Z, et al. Knowledge adaptation for efficient semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 578-587.

[80] Fang G, Song J, Shen C, et al. Data-free adversarial distillation. arXiv preprint arXiv:1912.11006, 2019.

[81] Chen Y, Li W, Van Gool L. Road: Reality oriented adaptation for semantic segmentation of urban scenes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7892-7901.

[82] Dou Q, Liu Q, Heng P A, et al. Unpaired multi-modal segmentation via knowledge distillation. IEEE transactions on medical imaging, 2020, 39(7):2415-2425.

[83] Tang S, Feng L, Shao W, et al. Learning efficient detector with semi-supervised adaptive distillation. arXiv preprint arXiv:1901.00366, 2019.

[84] Peng B, Jin X, Liu J, et al. Correlation congruence for knowledge distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 5007-5016.

[85] Tung F, Mori G. Similarity-preserving knowledge distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 1365-1374.

[86] Passalis N, Tefas A. Learning deep representations with probabilistic knowledge transfer//Proceedings of the 16th European Conference on Computer Vision. Munich, Germany, 2018: 268-284.

[87] Guan Y, Zhao P, Wang B, et al. Differentiable feature aggregation search for knowledge distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 469-484.

[88] Shen Z, He Z, Xue X. Meal: Multi-model ensemble via adversarial learning//Proceedings of the 33th AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 4886-4893.

[89] Ahn S, Hu S X, Damianou A, et al. Variational information distillation for knowledge transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9163-9171.

[90] Heo B, Lee M, Yun S, et al. Knowledge transfer via distillation of activation boundaries formed by hidden neurons//Proceedings of the 33th AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019:3779-3787.

[91] Liu I J, Peng J, Schwing A G. Knowledge flow: Improve upon your teachers. arXiv preprint arXiv:1904.05878, 2019.

[92] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780, 2017.

[93] Chen Z, Zhu L, Wan L, et al. A multi-task mean teacher for semisupervised shadow detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5611-5620.

[94] Chen Z, Dutton B, Ramachandra B, et al. Local clustering with mean teacher for semi-supervised learning//2020 25th International Conference on Pattern Recognition. Virtual Event, 2021:6243-6250.

[95] Fu S, Li Z, Xu J, et al. Interactive knowledge distillation. arXiv preprint arXiv:2007.01476, 2020.

[96] Shen C, Wang X, Yin Y, et al. Progressive network grafting for few-shot knowledge distillation. arXiv preprint arXiv:2012.04915, 2020.

[97] Park W, Kim D, Lu Y, et al. Relational knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3967-3976.

[98] Liu Y, Cao J, Li B, et al. Knowledge distillation via instance relationship graph//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7096-7104.

[99] Chen Z, Zheng X, Shen H, et al. Improving knowledge distillation via category structure// Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 205-219.

[100] Hou Y, Ma Z, Liu C, et al. Inter-region affinity distillation for road marking segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12486-12495.

[101] Tao X, Hong X, Chang X, et al. Few-shot class-incremental learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 12183-12192.

[102] Li Z, Jiang R, Aarabi P. Semantic relation preserving knowledge distillation for image-to-image translation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020:

648-663.

[103] Lee S, Song B C. Graph-based knowledge distillation by multi-head attention network. arXiv preprint arXiv:1907.02226, 2019.

[104] Ma J, Mei Q. Graph representation learning via multi-task knowledge distillation. arXiv preprint arXiv:1911.05700, 2019.

[105] Zhang C, Peng Y. Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 1135-1141.

[106] Lassance C, Bontonou M, Hacene G B, et al. Deep geometric knowledge distillation with graphs//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain, 2020: 8484-8488.

[107] Zhang W, Miao X, Shao Y, et al. Reliable data distillation on graph convolutional network//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. Portland, USA, 2020: 1399-1414.

[108] Zhan K, Niu C. Mutual teaching for graph convolutional networks. Future Generation Computer Systems, 2021, 115:837-843.

[109] Chen Y, Bian Y, Xiao X, et al. On self-distilling graph neural network. arXiv preprint arXiv:2011.02255, 2020.

[110] Zhang H, Lin S, Liu W, et al. Iterative graph self-distillation. arXiv preprint arXiv:2010.12609, 2020.

[111] Antaris S, Rafailidis D. Distill2vec: Dynamic graph representation learning with knowledge distillation//2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Hague, Netherlands, 2020: 60-64.

[112] Yang Y, Qiu J, Song M, et al. Distilling knowledge from graph convolutional networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7074-7083.

[113] Zhang Y, Xiang T, Hospedales T M, et al. Deep mutual learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4320-4328.

[114] Che D, Mei J P, Wang C, et al. Online knowledge distillation with diverse peers//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York City, USA, 2020: 3430-3437.

[115] Xue Q, Zhang W, Zha H. Improving domain-adapted sentiment classification by deep adversarial mutual learning//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York City, USA, 2020:9362-9369.

[116] Lan X, Zhu X, Gong S. Knowledge distillation by on-the-fly native ensemble//Proceedings of the 32th International Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 7528-7538.

[117] Xie J, Lin S, Zhang Y, et al. Training convolutional neural networks with cheap convolutions and online distillation. arXiv preprint arXiv:1909.13063, 2019.

[118] Wang L, Li D, Zhu Y, et al. Dual super-resolution learning for semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,USA, 2020: 3774-3783.

[119] Xiang L, Ding G, Han J. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 247-263.

[120] Song G, Chai W. Collaborative learning for deep neural networks. Advances in Neural Information Processing Systems, 2018, 31:1832-1841.

[121] Guo Q, Wang X, Wu Y, et al. Online knowledge distillation via collaborative learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,USA, 2020: 11020-11029.

[122] Guan S, Tai Y, Ni B, et al. Collaborative learning for faster stylegan embedding. arXiv preprint arXiv:2007.01758, 2020.

[123] Huang Z, Zou Y, Bhagavatula V, et al. Comprehensive attention self-distillation for weakly-supervised object detection. arXiv preprint arXiv:2010.12023, 2020.

[124] Liu B, Rao Y, Lu J, et al. Metadistiller: Network self-boosting via meta-learned top-down distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 694-709.

[125] Han S, Mao H, Dally W J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149, 2015.

[126] Zhang Z, Sabuncu M R. Self-distillation as instance-specific label smoothing. arXiv preprint arXiv:2006.05065, 2020.

[127] Lopes R G, Fenu S, Starner T. Data-free knowledge distillation for deep neural networks. arXiv preprint arXiv:1710.07535, 2017.

[128] Yin H, Molchanov P, Alvarez J M, et al. Dreaming to distill: Data-free knowledge transfer via deepinversion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 8715-8724.

[129] Micaelli P, Storkey A J. Zero-shot knowledge transfer via adversarial belief matching. Advances in Neural Information Processing Systems, 2019, 32:9551-9561.

[130] Chen H, Wang Y, Xu C, et al. Data-free learning of student networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 3514-3522.

[131] Nayak G K, Mopuri K R, Shaj V, et al. Zero-shot knowledge distillation in deep networks//Proceedings of the 36th International Conference on Machine Learning, Long Beach, USA, 2019: 4743-4751.

[132] Bhardwaj K, Suda N, Marculescu R. Dream distillation: A data-independent model compression framework. arXiv preprint arXiv:1905.07072, 2019.

[133] Park S, Kwak N. Feed: Feature-level ensemble for knowledge distillation. arXiv preprint arXiv:1909.10754, 2019.

[134] Walawalkar D, Shen Z, Savvides M. Online ensemble model

compression using knowledge distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 18-35.

[135] Du S, You S, Li X, et al. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. Advances in Neural Information Processing Systems, 2020, 33.

[136] Vapnik V, Izmailov R. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. Journal of Machine Learning Research, 2015, 16: 2023-2049.

[137] Tang F, Xiao C, Wang F, et al. Retaining privileged information for multi-task learning//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA, 2019: 1369-1377.

[138] Wang X, Zhang R, Sun Y, et al. Kdgan: Knowledge distillation with generative adversarial networks//Proceedings of the 32th International Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 783-794.

[139] Bhardwaj S, Srinivasan M, Khapra M M. Efficient video classification using fewer frames//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 354-363.

[140] Zhang F, Zhu X, Ye M. Fast human pose estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3517-3526.

[141] Buciluǎ C , Caruana R, Niculescu-Mizil A. Model compression//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, USA, 2006: 535-541.

[142] Srinivas S, Babu R V. Data-free parameter pruning for deep neural networks. arXiv preprint arXiv:1507.06149, 2015.

[143] Gupta S, Agrawal A, Gopalakrishnan K, et al. Deep learning with limited numerical precision//Proceedings of the 32th International Conference on Machine Learning, Lille, France, 2015: 1737-1746.

[144] Courbariaux M, Bengio Y, David J P. Binaryconnect: Training deep neural networks with binary weights during propagations//Advances in Neural Information Processing Systems. 2015: 3123-3131.

[145] Courbariaux M, Hubara I, Soudry D, et al. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. arXiv preprint arXiv:1602.02830, 2016.

[146] Rastegari M, Ordonez V, Redmon J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks//Proceedings of the 16th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 525-542.

[147] Mehta S, Rangwala H, Ramakrishnan N. Low rank factorization for compact multi-head self-attention. arXiv preprint arXiv:1912.00835, 2019.

[148] Lu Z, Sindhwani V, Sainath T N. Learning compact recurrent neural networks//2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China, 2016: 5960-5964.

[149] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

[150] Huang G, Liu S, Van Der Maaten L, et al. Condensenet: An efficient densenet using learned group convolutions//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA, 2018: 2752-2761.

[151] Iandola F N, Han S, Moskewicz M W, et al. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360, 2016.

[152] Ashok A, Rhinehart N, Beainy F, et al. N2n learning: Network to network compression via policy gradient reinforcement learning. arXiv preprint arXiv:1709.06030, 2017.

[153] Gao W, Wei Y, Li Q, et al. Pruning with hints: An efficient framework for model acceleration. Computer Science, 2018, 1-14.

[154] Miles R, Mikolajczyk K. Cascaded channel pruning using hierarchical self-distillation. arXiv preprint arXiv:2008.06814, 2020.

[155] Chen S, Wang W, Pan S J. Cooperative pruning in cross-domain deep neural network compression//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 2102-2108.

[156] Luo J H, Wu J. Neural network pruning with residual-connections and limited-data//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1458-1467.

[157] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017.

[158] Zhuang B, Shen C, Tan M, et al. Towards effective low-bitwidth convolutional neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7920-7928.

[159] Kim J, Bhalgat Y, Lee J, et al. Qkd: Quantization-aware knowledge distillation. arXiv preprint arXiv:1911.12491, 2019.

[160] Zhuang B, Liu L, Tan M, et al. Training quantized neural networks with a full-precision auxiliary module//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1488-1497.

[161] Ye J, Wang J, Zhang S. Distillation-guided residual learning for binary convolutional neural networks. arXiv preprint arXiv:2007.05223, 2020.

[162] Xu S, Li H, Zhuang B, et al. Generative low-bitwidth data free quantization//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 1-17.

[163] Cai Y, Yao Z, Dong Z, et al. Zeroq: A novel zero shot quantization framework//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13169-13178.

[164] Liu Y, Zhang W, Wang J. Zero-shot adversarial quantization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 1512-1521.

[165] Lin S, Ji R, Chen C, et al. Holistic cnn compression via low-rank

decomposition with knowledge transfer. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(12):2889-2905.

[166] Wang Z, Deng Z, Wang S. Accelerating convolutional neural networks with dominant convolutional kernel and knowledge preregression//Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 533-548.

[167] Zhao M, Li T, Abu Alsheikh M, et al. Through-wall human pose estimation using radio signals//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7356-7365.

[168] Thoker F M, Gall J. Cross-modal knowledge distillation for action recognition//2019 IEEE International Conference on Image Processing. Taipei, China, 2019: 6-10.

[169] Gupta S, Hoffman J, Malik J. Cross modal distillation for supervision transfer//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA, 2016: 2827-2836.

[170] Liu Y, Zhang W, Wang J. Source-free domain adaptation for semantic segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Event, 2021: 1215-1224.

[171] Michieli U, Zanuttigh P. Knowledge distillation for incremental learning in semantic segmentation. Computer Vision and Image Understanding, 2021, 205:103167.

[172] Zhai M, Chen L, Tung F, et al. Lifelong gan: Continual learning for conditional image generation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 2759-2768.

[173] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, USA, 2017: 4681-4690.

[174] Wang X, Yu K, Wu S, et al. Esrgan: Enhanced super-resolution generative adversarial networks//Proceedings of the 15th European Conference on Computer Vision workshops. Munich, Germany, 2018: 63-79.

[175] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4401-4410.

[176] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096, 2018.

[177] Park T, Liu M Y, Wang T C, et al. Semantic image synthesis with spatially-adaptive normalization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2337-2346.

[178] Karnewar A, Wang O. Msg-gan: multi-scale gradient gan for stable image synthesis. arXiv preprint arXiv:1903.06048, 2019.

[179] Aguinaldo A, Chiang P Y, Gain A, et al. Compressing gans using knowledge distillation. arXiv preprint arXiv:1902.00159, 2019.

[180] Chen H, Wang Y, Shu H, et al. Distilling portable generative adversarial networks for image translation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York City, USA, 2020: 3585-3592.

[181] Goldblum M, Fowl L, Feizi S, et al. Adversarially robust distillation//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York City, USA, 2020: 3996-4003.

[182] Chung I, Park S, Kim J, et al. Feature-map-level online adversarial knowledge distillation//Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020: 2006-2015.

[183] Wang W, Hong W, Wang F, et al. Gan-knowledge distillation for one-stage object detection. IEEE Access, 2020, 8:60719-60727.

[184] Heo B, Lee M, Yun S, et al. Knowledge distillation with adversarial samples supporting decision boundary//Proceedings of the 33th AAAI Conference on Artificial Intelligence. Honolulu,USA, 2019: 3771-3778.

[185] Sutton R S, Barto A G. Reinforcement learning: An introduction. MIT press, 2018.

[186] Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 2016, 17(1):1334-1373.

[187] Silver D, Schritiwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. Nature, 2017, 550(7676): 354-359.

[188] Zha D, Lai K H, Cao Y, et al. Rlcard: A toolkit for reinforcement learning in card games. arXiv preprint arXiv:1910.04376, 2019.

[189] Hong Z W, Nagarajan P, Maeda G. Periodic intra-ensemble knowledge distillation for reinforcement learning//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Bilbao, Spain, 2021: 87-103.

[190] Lai K H, Zha D, Li Y, et al. Dual policy distillation. arXiv preprint arXiv:2006.04061, 2020.

[191] Burda Y, Edwards H, Storkey A, et al. Exploration by random network distillation. arXiv preprint arXiv:1810.12894, 2018.

[192] Xue Z, Luo S, Wu C, et al. Transfer heterogeneous knowledge among peer-to-peer teammates: A model distillation approach. arXiv preprint arXiv:2002.02202, 2020.

[193] Gao Z, Xu K, Ding B, et al. Knowru: Knowledge reusing via knowledge distillation in multi-agent reinforcement learning. arXiv preprint arXiv:2103.14891, 2021.

[194] Wadhwania S, Kim D K, Omidshafiei S, et al. Policy distillation and value matching in multiagent reinforcement learning//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China, 2019: 8193-8200.

[195] Liu Q, Xie L, Wang H, et al. Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 3662-3671.

[196] Gu J, Tresp V. Search for better students to learn distilled

knowledge//Proceedings of the 24th European Conference on Artificial Intelligence. Santiago de Compostela, Spain, 2020: 1159-1165.

[197] Zhou W, Xu C, Mcauley J. Meta learning for knowledge distillation. arXiv preprint arXiv:2106.04570, 2021.

[198] Flennerhag S, Moreno P G, Lawrence N D, et al. Transferring knowledge across learning processes. arXiv preprint arXiv:1812.01054, 2018.

[199] Jang Y, Lee H, Hwang S J, et al. Learning what and where to transfer//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 3030-3039.

[200] Yao Q, Wang M, Chen Y, et al. Taking human out of learning applications: A survey on automated machine learning. arXiv preprint arXiv:1810.13306, 2018.

[201] Liu Y, Jia X, Tan M, et al. Search to distill: Pearls are everywhere but not the eyes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7539-7548.

[202] Wei L, Xiao A, Xie L, et al. Circumventing outliers of autoaugment with knowledge distillation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 608-625.

[203] Macko V, Weill C, Mazzawi H, et al. Improving neural architecture search image classifiers via ensemble learning. arXiv preprint arXiv:1903.06236, 2019.

[204] Li X, Lin C, Li C, et al. Improving one-shot nas by suppressing the posterior fading//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13836-13845.

[205] Wu B, Dai X, Zhang P, et al. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 10734-10742.

[206] Li C, Peng J, Yuan L, et al. Block-wisely supervised neural architecture search with knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1989-1998.

[207] Dong X, Yang Y. Network pruning via transformable architecture search. arXiv preprint arXiv:1905.09717, 2019.

[208] Ge S, Zhao S, Li C, et al. Low-resolution face recognition in the wild via selective knowledge distillation. IEEE Transactions on Image Processing, 2018, 28(4):2051-2062.

[209] Kong H, Zhao J, Tu X, et al. Cross-resolution face recognition via prior-aided face hallucination and residual knowledge distillation. arXiv preprint arXiv:1905.10777, 2019.

[210] Yan M, Zhao M, Xu Z, et al. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition//Proceedings of the IEEE/CVF IEEE/CVF International Conference on Computer Vision Workshop. Seoul, Korea, 2019: 2647-2654.

[211] Karlekar J, Feng J, Wong Z S, et al. Deep face recognition model compression via knowledge transfer and distillation. arXiv preprint arXiv:1906.00619, 2019.

[212] Wu X, He R, Hu Y, et al. Learning an evolutionary embedding via massive knowledge distillation. International Journal of Computer Vision, 2020, 128(8):2089-2106.

[213] Shen J, Vesdapunt N, Boddeti V N, et al. In teacher we trust: Learning compressed models for pedestrian detection. arXiv preprint arXiv:1612.00478, 2016.

[214] Kruthiventi S S S, Sahay P, Biswal R. Low-light pedestrian detection from rgb images using multi-modal knowledge distillation//2017 IEEE International Conference on Image Processing. Beijing, China, 2017: 4207-4211.

[215] Chen R, Ai H, Shang C, et al. Learning lightweight pedestrian detector with hierarchical knowledge distillation//2019 IEEE International Conference on Image Processing. Taipei, China, 2019: 1645-1649.

[216] Martínez-González A, Villamizar M, Canévet O, et al. Efficient convolutional neural networks for depth-based multi-person pose estimation. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(11):4207-4221.

[217] Nie X, Li Y, Luo L, et al. Dynamic kernel distillation for efficient pose estimation in videos//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 6942-6950.

[218] Wang C, Kong C, Lucey S. Distill knowledge from nrsfm for weakly supervised 3d pose learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 743-752.

[219] Hwang D H, Kim S, Monet N, et al. Lightweight 3d human pose estimation network training using teacher-student learning//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Snowmass Village, USA, 2020: 479-488.

[220] Zhang S, Guo S, Wang L, et al. Knowledge integration networks for action recognition//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York City, NY, USA, 2020: 12862-12869.

[221] Tang R, Lu Y, Liu L, et al. Distilling task-specific knowledge from bert into simple neural networks. arXiv preprint arXiv:1903.12136, 2019.

[222] Sun S, Cheng Y, Gan Z, et al. Patient knowledge distillation for bert model compressions. arXiv preprint arXiv:1908.09355, 2019.

[223] Sanh V, Debut L, Chaumond J, et al. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

[224] Jiao X, Yin Y, Shang L, et al. Tinybert: Distilling bert for natural language understanding//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Virtual Event, 2020:4163-4174.

[225] Sun Z, Yu H, Song X, et al. Mobilebert: Task-agnostic compression of bert by progressive knowledge transfer//Proceedings of the 58th

Annual Meeting of the Association for Computational Linguistics. Virtual Event, 2020: 2158–2170

[226] Zhou C, Neubig G, Gu J. Understanding knowledge distillation in non-autoregressive machine translations. arXiv preprint arXiv:1911.02727, 2019.

[227] Tan X, Ren Y, He D, et al. Multilingual neural machine translation with knowledge distillation. arXiv preprint arXiv:1902.10461, 2019.

[228] Wei H R, Huang S, Wang R, et al. Online distilling from checkpoints for neural machine translation//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA, 2019: 1932-1941.

[229] Gordon M A, Duh K. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. arXiv preprint arXiv:1912.03334, 2019.

[230] Wang W, Zhang J, Zhang H, et al. A teacher-student framework for maintainable dialog manager//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018:3803-3812.

[231] Hu M, Peng Y, Wei F, et al. Attention-guided answer distillation for machine reading comprehension. arXiv preprint arXiv:1808.07644, 2018.

[232] Arora S, Khapra M M, Ramaswamy H G. On knowledge distillation from complex networks for response prediction//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, USA, 2019: 3813-3822.

[233] Yang Z, Shou L, Gong M, et al. Model compression with two stage multi-teacher knowledge distillation for web question answering system//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, USA, 2020: 690-698.

[234] Mukherjee S, Awadallah A H. Distilling bert into simple neural networks with unlabeled transfer data. arXiv preprint arXiv:1910.01769, 2019.

[235] Zhao S, Gupta R, Song Y, et al. Extremely Small BERT Models from Mixed-Vocabulary Training. arXiv preprint arXiv: 1909.11687, 2019.

[236] Xu Y, Qiu X, Zhou L, et al. Improving bert fine-tuning via self-ensemble and self-distillation. arXiv preprint arXiv:2002.10345, 2020.

[237] Liu W, Zhou P, Wang Z, et al. Fastbert: a self-distilling bert with adaptive inference time//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Virtual Event, 2020: 6035-6044.

[238] Zhou G, Fan Y, Cui R, et al. Rocket launching: A universal and efficient framework for training well-performing light net//Proceedings of the 32th AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018.

[239] Xu C, Li Q, Ge J, et al. Privileged features distillation at taobao recommendations//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Virtual Event, 2020: 2590-2598.

[240] Mi F, Lin X, Faltings B. Ader: Adaptively distilled exemplar replay towards continual learning for session-based recommendation//Proceedings of the 14th ACM Conference on Recommender Systems. Virtual Event, 2020: 408-413.

[241] Zhu J, Liu J, Li W, et al. Ensembled ctr prediction via knowledge distillation//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Virtual Event, Ireland, 2020: 2941-2958.

[242] Kang S, Hwang J, Kweon W, et al. De-rrd: A knowledge distillation framework for recommender system//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Galway, Ireland, 2020: 605-614.

[243] Watanabe S, Hori T, Le Roux J, et al. Student-teacher network learning with enhanced features//2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, USA, 2017: 5275-5279.

[244] Oord A, Li Y, Babuschkin I, et al. Parallel wavenet: Fast high-fidelity speech synthesis//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden, 2018: 3918-3926.

[245] Gao L, Mi H, Zhu B, et al. An adversarial feature distillation method for audio classification. IEEE Access, 2019, 7:105319-105330.

[246] Shen P, Lu X, Li S, et al. Interactive learning of teacher-student model for short utterance spoken language identification//2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK, 2019: 5981-5985.

[247] Perez A, Sanguineti V, Morerio P, et al. Audio-visual model distillation using acoustic images//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Snowmass Village, USA, 2020: 2854-2863.

[248] Zmora N, Jacob G, Zlotnik L, et al. Neural network distiller: A python package for dnn compression research. arXiv preprint arXiv:1910.12232, 2019.

[249] Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 2019, 32:8026-8037.

[250] Yang Z, Cui Y, Chen Z, et al. Textbrewer: An open-source knowledge distillation toolkit for natural language processing. arXiv preprint arXiv:2002.12620, 2020.

[251] MATSUBARA Y. torchdistill: A modular, configuration-driven framework for knowledge distillation//International Workshop on Reproducible Research in Pattern Recognition. Virtual Event, 2021: 24-44.

[252] Shah H, Khare A, Shah N, et al. Kd-lib: A pytorch library for knowledge distillation, pruning and quantization. arXiv preprint arXiv:2011.14691, 2020.

[253] Cheng Keyang, Wang Ning, Shi Wenxi, Zhan Yongzhao. Research Advances in the Interpretability of Deep Learning. Journal of Computer Research and Development, 2020, 57(6): 1208-1217. (in Chinese)

(成科扬, 王宁, 师文喜, 等. 深度学习可解释性研究进展. 计算机研究与发展, 2020, 57(6):10.)

[254] Hua Yingying, Zhang Daichi, Ge Shiming. Research Progress in the Interpretability of Deep Learning Models. Journal of Cyber Security. 2020, 5(3):12. (in Chinese)
(化盈盈, 张岱墀, 葛仕明. 深度学习模型可解释性的研究进展. 信息安全学报, 2020, 5(3):12.)

[255] Schindler K. An overview and comparison of smooth labeling methods for land-cover classification. IEEE transactions on geoscience and remote sensing, 2012, 50(11):4534-4545.

[256] Tang J, Shivanna R, Zhao Z, et al. Understanding and improving knowledge distillation. arXiv preprint arXiv:2002.03532, 2020.

[257] Liang R, Li T, Li L, et al. Knowledge consistency between neural networks and beyond. arXiv preprint arXiv:1908.01581, 2019.

[258] Cho J H, Hariharan B. On the efficacy of knowledge distillation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 4794-4802.

[259] Cheng X, Rao Z, Chen Y, et al. Explaining knowledge distillation by quantifying the knowledge//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA 2020: 12925-12935.

[260] Rahbar A, Panahi A, Bhattacharyya C, et al. On the unreasonable effectiveness of knowledge distillation: Analysis in the kernel regime. arXiv preprint arXiv:2003.13438, 2020.

[261] Menon A K, Rawat A S, Reddi S J, et al. Why distillation helps: a statistical perspective. arXiv preprint arXiv:2005.10419, 2020.

[262] Jha N K, Saini R, Mittal S. On the demystification of knowledge distillation: A residual network perspective. arXiv preprint arXiv:2006.16589, 2020.

[263] Ji G, Zhu Z. Knowledge distillation in wide neural networks: Risk bound, data efficiency and imperfect teacher. arXiv preprint arXiv:2010.10090, 2020.

[264] Saglietti L, Zdeborová L. Solvable model for inheriting the regularization through knowledge distillation. arXiv preprint arXiv:2012.00194, 2020.

[265] Mobahi H, Farajtabar M, Bartlett P L. Self-distillation amplifies regularization in hilbert space. arXiv preprint arXiv:2002.05715, 2020.

[266] Hsu D, Ji Z, Telgarsky M, et al. Generalization bounds via distillation. arXiv preprint arXiv:2104.05641, 2021.

[267] Gong C, Chang X, Fang M, et al. Teaching semi-supervised classifier via generalized distillation//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 2156-2162.

[268] Zhu Z, Hong J, Zhou J. Data-free knowledge distillation for heterogeneous federated learning. arXiv preprint arXiv:2105.10056, 2021.

[269] Fu Y, Chem W, Wang H, et al. Autogan-distiller: Searching to compress generative adversarial networks//Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 2020: 3292-3303.

[270] Jaiswal A, Abdalmageed W, Wu Y, et al. Capsulegan: Generative adversarial capsule network//Proceedings of the 16th European Conference on Computer Vision. Munich, Germany, 2018: 526-535.

[271] Xiang C, Zhang L, Tang Y, et al. Ms-capsnet: A novel multi-scale capsule network. IEEE Signal Processing Letters, 2018, 25 (12):1850-1854.

[272] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[273] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK, 2020: 213-229.

[274] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.

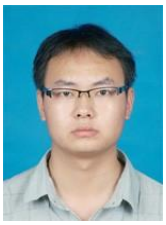
[275] Pearl J, Mackenzie D. The book of why: the new science of cause and effect. Basic books, 2018.

[276] Louizos C, Shalit U, Mooij J, et al. Causal effect inference with deep latent-variable models//Proceedings of the 31th International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 6449-6459.

LIU Yu-Ang, Ph.D. candidate. His research interestsinclude knowledge distillation, model compression and computer vision.

ZHANG Wei, Ph.D., associate professor. His research interests include data mining, recommender system.

WANG Jun, Ph.D., professor. His research interests include machine learning, computer vision.



SHAO Ren-Rong, Ph.D. candidate. His research interests include computer vision, and model compression.

Background

With the rapid development of deep learning, it also faces a lot of challenges. Firstly, the model has become more and more complex, making it unable to be deployed on embedded devices and mobile devices. Secondly, deep learning seriously relies on manual labels, causing huge costs and expenses. Finally, the security and privacy issues of deep learning have gradually attracted researchers' attention. Knowledge distillation is one of the important technologies of model compression. It belongs to the category of transfer learning, it can not only be used for model compression but also integrate with various fields, and can solve many difficulties of deep learning at present. Therefore, knowledge distillation has gradually been valued by researchers since it was proposed. Because of its simplicity, ease of use, and scalability, it has been widely concerned and applied in the industry at present.

Our team has been committed to research in areas such as model compression and knowledge distillation and has achieved certain research results in data-free distillation and model quantification. Due to the rapid development of knowledge distillation and the numerous and complicated articles, it is difficult for beginners to get a complete picture. During the research, our team has collected a lot of materials and documents, but most of them are in English, which is not convenient for domestic researchers to read and learn. To introduce the research status and solution of knowledge distillation and to facilitate researchers to understand knowledge distillation from a macro perspective.

We categorized and summarized based on current research, introduced the current research status, as well as some representative methods and ideas in the process of integrating with various fields, and discussed possible development trends of the future. We published and shared this article in Chinese to facilitate more domestic researchers to learn. At the same time, we also hope to attract more researchers' attention and contribute in this direction.