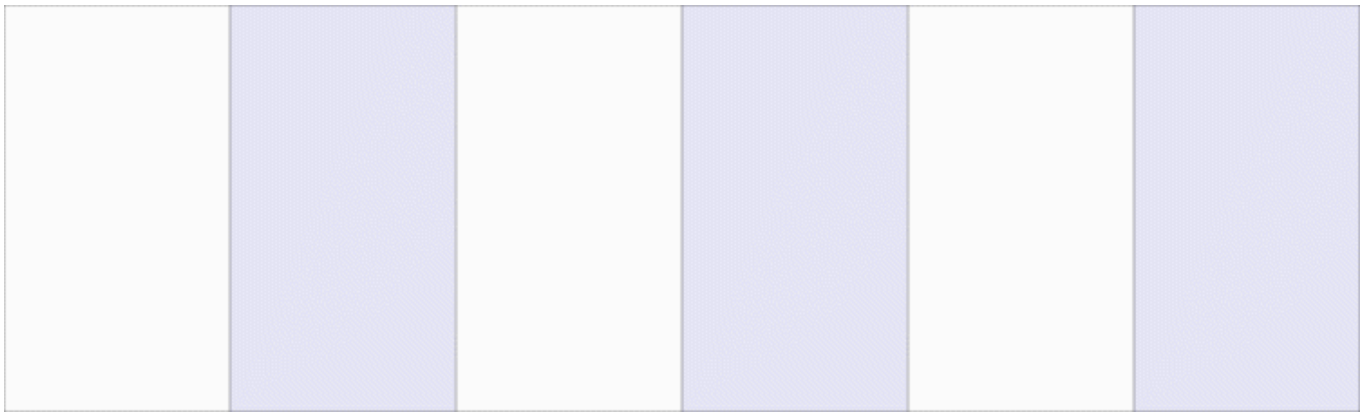
「深度学习中知识蒸馏」研究综述

CV开发者都爱看的 极市平台 2023-02-06 22:00:48 发表于广东 手机阅读



来源丨专知

编辑丨极市平台

极市导读

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，文末附PDF下载。>>加入极市CV技术交流群，走在计算机视觉的最前沿

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得深而复杂，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**.

第 45 卷 第 8 期 计 算 机 学 报 Vol. 45 No. 8
2022 年 8 月 CHINESE JOURNAL OF COMPUTERS Aug. 2022

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）

随着深度神经网络的崛起和演化，深度学习在计算机视觉、自然语言处理、推荐系统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得深而复杂：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；伴随着模型的复杂化，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，其在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等人于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．到了发展期，知识的形式逐渐丰富、多元，不再局限于单一的节点，这一时期较为代表性的蒸馏方法有AT、FT．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、跨领域、持续学习、隐私保护等；在和

其他领域交叉过程中又逐渐结合了对抗学习、强化学习、元学习、自动机器学习、自监督学习等．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．

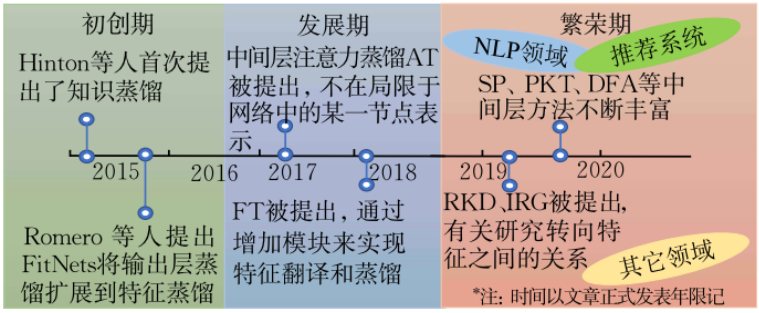


图 1　知识蒸馏发展历程

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着这一些挑战；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，均发表于2021年．相较于前者，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，完整地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 节首先介绍知识蒸馏的理论基础及分类；第 3 ～ 6 节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分类和对比，并分析不同研究方向面临的机遇和挑战；第 7 节列举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 节对知识蒸馏的原理和可解释性方面的工作进行梳理；最后，对知识蒸馏在深度学习场景下的未来发展趋势提出一些见解，并进行全文总结．

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他模型的训练，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”．

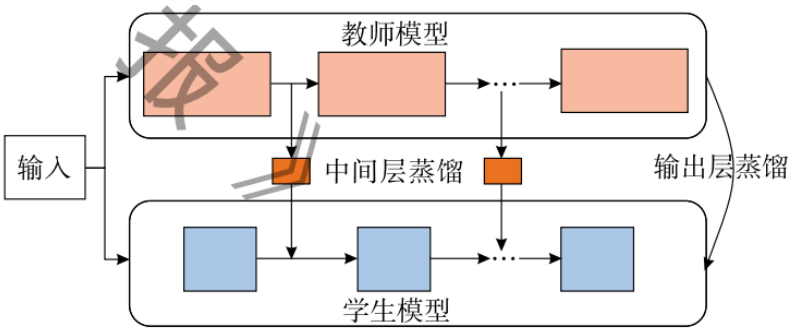


图 3　知识蒸馏教师学生模型结构流程图

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.

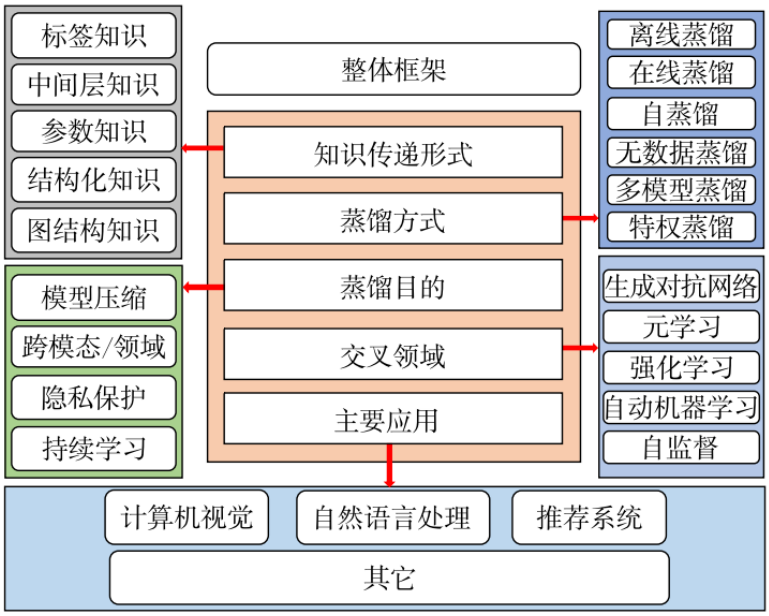


图 4　知识蒸馏整体分类框架

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	$\mathcal{L}_{\mathrm{KD}}=\mathcal{H}(\boldsymbol{y}_{\mathrm{true}},P_S)+\lambda\mathcal{H}(P_T^{\tau},P_S^{\tau})$	$\mathcal{H}$ 为交叉熵，λ 为平衡因子. $\boldsymbol{y}_{\mathrm{true}}$ 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	$\mathcal{L}_{\mathrm{KD}}=\Vert f_T(\boldsymbol{x};\boldsymbol{W}_T)-f_S(\boldsymbol{x};\boldsymbol{W}_S)\Vert_F$	**同构蒸馏**：f_T 和 f_S 分别为教师和学生模型，$\boldsymbol{x}$ 为中间层输入的特征；$\boldsymbol{W}_T$ 和 $\boldsymbol{W}_S$ 为教师和学生模型的中间层参数，F 为范数值.
	$\mathcal{L}_{\mathrm{KD}}=\Vert f_S(\boldsymbol{x};\boldsymbol{W}_S)-\gamma(f_T(\boldsymbol{x};\boldsymbol{W}_T),\boldsymbol{W}_R)\Vert_F$	**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，$\boldsymbol{W}_R$ 为适配器训练参数.
参数知识	$\mathcal{L}=\mathop{\mathbb{E}}_{\{\boldsymbol{x}_i\}_{i=1}^{n_t+n_u}} D(f(\boldsymbol{x}_i,\xi';\theta'),f(\boldsymbol{x}_i,\xi'';\theta''))$ 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.
	$\mathcal{L}(\boldsymbol{Q}_j)=\sum_{j=1}^{M}\sum_{i=1}^{N}\Vert\boldsymbol{Q}_j*\mathrm{Y}_{ij}^S-\mathrm{Y}_{ij}^T\Vert_F$ 其中，$Y^S,Y^T\in\mathbb{R}^{n_c\times d}$	**模块注入**：$\boldsymbol{Q}_j$ 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	$\mathcal{L}_{\mathrm{SKD}}=\sum_{(x_i,\cdots,x_n)\in\mathcal{X}^n} D(\psi(t_1,\cdots,t_n),\psi(s_1,\cdots,s_n))$	其中，$(x_1,\cdots,x_n)$ 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	$\mathcal{L}_{\mathrm{GKD}}=\sum_{\ell\in\Lambda}D(\mathcal{G}_\ell^S(\boldsymbol{x}),\mathcal{G}_\ell^T(\boldsymbol{x}))$ 其中，$\mathcal{G}_\ell(\boldsymbol{x})=\langle\boldsymbol{x}_\ell,\boldsymbol{W}_\ell\rangle$	$\mathcal{G}_\ell^S$ 和 $\mathcal{G}_\ell^T$ 为学生和教师在 ℓ 层上的图结构构造函数，$\boldsymbol{x}_\ell$ 为输入节点，$\boldsymbol{W}_\ell$ 表示边的权重矩阵.

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 l o g i t s 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，后面介绍的各类蒸馏方法或具体应用都是以此为基础. 相关的优缺点和实验对比，见表 2 ～表 4 所示.

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

表 3 不同“知识”表达形式的优缺点

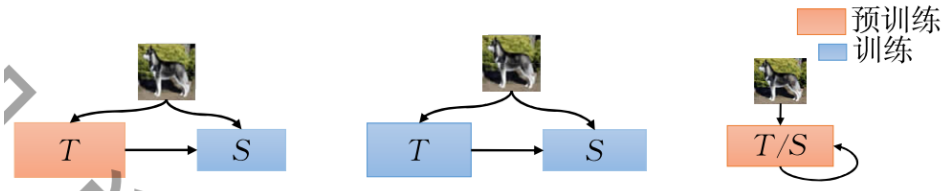
知识形式	优缺点	解决的问题	适用场景
标签知识	优点：方法简单通用，易于实现. 缺点：知识单一，依赖于损失函数的设计且对参数敏感.	较为通用，一般作为各种任务中知识蒸馏的基础.	适合分类，识别，分割等几乎所有任务.
中间层知识	优点：具有丰富知识信息，能够满足复杂任务对知识的需求. 缺点：知识形式多样，无法有效的整合且互相影响，试错代价高.	解决了输出层知识信息单一，不够丰富的不足，能够为模型提供较为丰富的特征知识.	适用于安全隐私要求相对不高，教师模型可访问的场景；对特征依赖较大场景且模型准确率有较高要求的情况.
参数知识	优点：模型训练稳定且效率较高，设计精巧. 缺点：较难实现，不利于端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.

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.

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．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．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，如图 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；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，）．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，形成特权蒸馏学习．接下来，将分别介绍不同蒸馏学习方式的代表性工作．



（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）

主要应用

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，而对于其他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，视觉中还有一些其他应用比如视频分类［105］、深度估计和光流 / 场景流估计［169］等等．

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		√			√		√	√			√	√		

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, N MT），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于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．

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			✓				✓	✓			✓			

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 DistillB

ERT，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。

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．

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	✓						✓	✓			✓			

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．

公众号后台回复“知识蒸馏”获取知识蒸馏研究综述PDF