软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(2):675−710 [doi: 10.13328/j.cnki.jos.006933]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

图知识蒸馏综述: 算法分类与应用分析*

刘 静[1,2], 郑铜亚[3], 郝沁汾[1]

[1](处理器芯片全国重点实验室 (中国科学院 计算技术研究所), 北京 100190)

[2](中国科学院大学 计算机科学与技术学院, 北京 100049)

[3](浙江大学 计算机科学与技术学院, 浙江 杭州 310027)

通信作者: 郝沁汾, E-mail: haoqinfen@ict.ac.cn

摘 要: 图数据, 如引文网络, 社交网络和交通网络, 广泛地存在现实生活中. 图神经网络凭借强大的表现力受到广泛关注, 在各种各样的图分析应用中表现卓越. 然而, 图神经网络的卓越性能得益于标签数据和复杂的网络模型, 而标签数据获取困难且计算资源代价高昂. 为了解决数据标签的稀疏性和模型计算的高复杂性问题, 知识蒸馏被引入到图神经网络中. 知识蒸馏是一种利用性能更好的大模型 (教师模型) 的软标签监督信息来训练构建的小模型 (学生模型), 以期达到更好的性能和精度. 因此, 如何面向图数据应用知识蒸馏技术成为重大研究挑战, 但目前尚缺乏对于图知识蒸馏研究的综述. 旨在对面向图的知识蒸馏进行全面综述, 首次系统地梳理现有工作, 弥补该领域缺乏综述的空白. 具体而言, 首先介绍图和知识蒸馏背景知识; 然后, 全面梳理 3 类图知识蒸馏方法, 面向深度神经网络的图知识蒸馏、面向图神经网络的图知识蒸馏和基于图知识的模型自蒸馏方法, 并对每类方法进一步划分为基于输出层、基于中间层和基于构造图知识方法; 随后, 分析比较各类图知识蒸馏算法的设计思路, 结合实验结果总结各类算法的优缺点; 此外, 还列举图知识蒸馏在计算机视觉、自然语言处理、推荐系统等领域的应用; 最后对图知识蒸馏的发展进行总结和展望. 还将整理的图知识蒸馏相关文献公开在 GitHub 平台上, 具体参见: https://github.com/liujing1023/Graph-based-Knowledge-Distillation.

关键词: 图数据; 图神经网络; 知识蒸馏

中图法分类号: TP18

中文引用格式: 刘静, 郑铜亚, 郝沁汾. 图知识蒸馏综述: 算法分类与应用分析. 软件学报, 2024, 35(2): 675–710. http://www.jos.org.cn/1000-9825/6933.htm

英文引用格式: Liu J, Zheng TY, Hao QF. Survey on Knowledge Distillation with Graph: Algorithms Classification and Application Analysis. Ruan Jian Xue Bao/Journal of Software, 2024, 35(2): 675–710 (in Chinese). http://www.jos.org.cn/1000-9825/6933.htm

Survey on Knowledge Distillation with Graph: Algorithms Classification and Application Analysis

LIU Jing[1,2], ZHENG Tong-Ya[3], HAO Qin-Fen[1]

[1](State Key Lab of Processors (Institute of Computing Technology, Chinese Academy of Sciences), Beijing 100190, China)

[2](School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China)

[3](College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

Abstract: Graph data, such as citation networks, social networks, and transportation networks, exist widely in the real world. Graph neural networks (GNNs) have attracted extensive attention due to their strong expressiveness and excellent performance in a variety of graph analysis applications. However, the excellent performance of GNNs benefits from label data which are difficult to obtain, and complex network models with high computational costs. Knowledge distillation (KD) is introduced into the GNNs to address the labeled data scarcity and high complexity of GNNs. KD is a method of training constructed small models (student models) by soft-label supervision information from larger models (teacher models) to yield better performance and accuracy. Therefore, how to apply the KD technology to

* 收稿时间: 2022-09-07; 修改时间: 2022-11-03, 2023-01-20; 采用时间: 2023-03-15; jos 在线出版时间: 2023-08-23
CNKI 网络首发时间: 2023-08-28

graph data has become a research challenge, but there is still a lack of a graph-based KD research review. Aiming at providing a comprehensive overview of KD based on graphs, this study first summarizes the existing studies and fills in the review gap in this field. Specifically, this study first introduces the background knowledge of graph and KD. Then, three types of graph-based knowledge distillation methods are comprehensively summarized, including graph knowledge distillation for deep neural networks (DNNs), graph knowledge distillation for GNNs, and self-KD-based graph knowledge distillation. Furthermore, each type of method is further divided into knowledge distillation methods based on the output layer, the middle layer, and the constructed graph. Subsequently, the design ideas of various graph-based knowledge distillation algorithms are analyzed and compared, and the advantages and disadvantages of the algorithms are concluded with experimental results. In addition, the application of graph-based knowledge distillation in computer vision, natural language processing, recommendation systems, and other fields are also listed. Finally, the development of graph-based knowledge distillation is summarized and prospected. This study also discloses the references related to graph-based knowledge distillation on GitHub. Please refer to https://github.com/liujing1023/Graph-based-Knowledge-Distillation.

Key words: graph data; graph neural network (GNN); knowledge distillation (KD)

图数据 (graph data)[1], 作为一种表示物体与物体之间关系的重要数据类型, 被广泛地应用在现实世界中任务场景中, 如用户推荐[2]、药物发现[3]、交通预测[4]、点云分类[5]和芯片设计[6]等. 不同于欧氏空间中的结构化数据, 图数据的结构复杂, 蕴含着丰富的信息. 为了从复杂的图数据中学习到包含充分信息的向量化表示, 越来越多的研究开始将深度学习方法应用到图数据领域. 借鉴卷积神经网络[7]的思想, 图神经网络 (GNN)[8]被提出. 实践证明, 图神经网络已经在节点分类[9]、链接预测[10]、图分类[11]等任务得到了有效应用.

然而, 随着卷积算子的完善和大图规模的发展, 研究者开始考虑如何训练出高精度且高效的图卷积神经网络, 如尝试训练更深的网络来增强模型的泛化能力. 众所周知, 图神经网络作为半监督学习的方法, 其卓越的性能严重依赖于大量的高质量标签数据和高度复杂的网络模型, 而数据标签获取困难且大规模图模型计算存储代价高昂.

为了有效缓解 GNN 面临的数据标签稀疏性和模型计算的高复杂性问题, 知识蒸馏[12]技术被引入到图分析研究中. 知识蒸馏是一种基于“教师-学生 (teacher-student, T-S)”网络思想的训练方法, 旨在通过将复杂、学习能力强的大网络模型 (教师模型) 学到的软标签知识蒸馏出来, 传递给参数量小、学习能力弱的小网络模型 (学生模型), 用以提高小网络模型的性能, 从而达到接近大网络模型的效果, 最终实现模型压缩的目的. 由于知识蒸馏简单有效, 在学术界和工业界也获得一系列显著的成功应用, 如计算机视觉[13]、语音识别[14]、自然语言处理[15]等.

近年来, “教师-学生”知识蒸馏框架在训练图神经网络方面显示出其潜力. 受到知识蒸馏在卷积神经网络上成功应用的驱动, 研究者尝试在图数据上或直接在图神经网络上设计知识蒸馏算法, 将知识蒸馏和图神经网络相结合. LSP[16]是第 1 个将知识蒸馏应用到图卷积神经网络 (GCN)[17]上的工作, 利用提出的局部结构保留模块将预训练深层 GCN 教师模型中的局部图结构知识蒸馏到具有较少参数的浅层 GCN 学生模型中. 随后, 也有一些其他面向图的知识蒸馏工作陆续被提出. 尽管知识蒸馏在图神经网络上取得了令人鼓舞的进展, 但现有的知识蒸馏方法一直集中在以结构化网格数据作为输入的卷积神经网络 (CNN) 上, 而在具有不规则数据处理能力的图神经网络上的研究却很少, 且目前缺乏对于图知识蒸馏研究的综述. 本文旨在对面向图的知识蒸馏进行全面综述, 首次系统地梳理现有知识蒸馏在图上的工作, 以弥补该领域缺乏综述的空白.

本文组织结构如后文图 1 所示, 本文第 1 节介绍图和知识蒸馏定义并给出图知识蒸馏研究的最新进展. 第 2 节梳理了面向图的知识蒸馏方法, 包括面向深度神经网络的图知识蒸馏、面向图神经网络的图知识蒸馏和基于图知识的模型自蒸馏 3 大类方法. 第 3、4、5 节按照蒸馏位置的不同, 针对不同类别的图知识蒸馏方法进一步细分为基于输出层、基于中间层和基于构造图的知识蒸馏方法, 并对现有的方法进行详细归纳和总结. 第 6 节对比分析经典的图知识蒸馏算法. 第 7 节列举图知识蒸馏方法在计算机视觉、自然语言处理、推荐系统等场景中的应用. 第 8 节展望面向图的知识蒸馏未来研究方向. 最后一节总结全文.

1 图知识蒸馏相关背景知识

在具体介绍图知识蒸馏方法之前, 本节先给出图知识蒸馏技术涉及的基本概念和符号定义, 接着对图知识蒸馏方法的发展和划分标准进行简要说明.

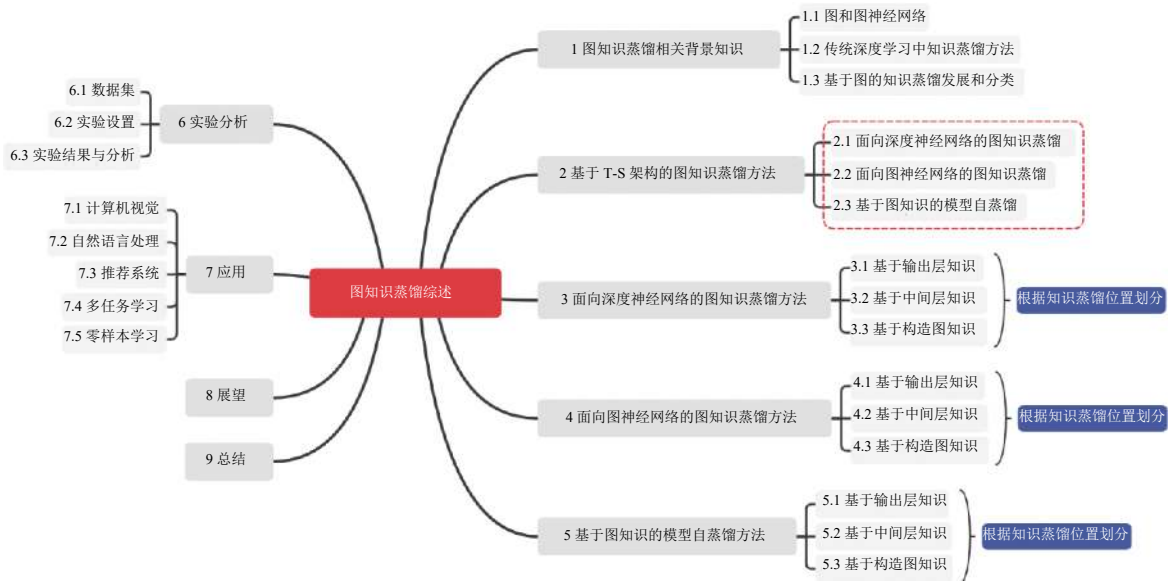


图 1　本文的组织结构图

1.1　图和图神经网络

尽管深度学习已经在欧氏空间的结构化数据中取得了很大的成功, 但现实生活中的数据一般自然的建模为图这样的非结构化数据. 图 (graph) 作为一种常见的数据结构, 可以表示为顶点 V 和边 E 的集合, 记为 $G=(V,E)$. 由于图结构的强大表现力, 其被广泛地应用于图分析研究中. 例如, 在电商推荐[18]领域, 需要一个基于图的学习系统, 能够利用用户和产品之间的交互实现高度精准的推荐. 图数据的复杂性对现有机器学习算法提出了重大挑战: 每张图大小不同、节点无序, 且一张图中的每个节点都有不同数目的邻居节点, 使得深度学习中常规的卷积运算无法直接应用于图数据. 近年来, 借鉴 CNN 的卷积思想, 研究者开始尝试将深度学习方法应用到图分析领域. 自此, 图神经网络 (GNN)[8]诞生, 由于其较好的性能和可解释性, GNN 最近已成为一种广泛应用的图分析方法.

最近, 深度学习领域关于图神经网络的研究热情日益高涨, 已经成为各大领域的研究热点. GNN 处理非结构化数据时的出色能力使其在生物化学[19]、物理建模[20]、知识图谱[21]和电路设计[22]等方面都取得了新的突破. 随着图神经网络模型的发展, GNN 主要可以分为谱方法和空间方法.

一方面, 基于谱的方法从图信号处理的角度引入滤波器来定义图卷积. 2013 年 Bruna 等人[23]基于谱理论[24]将频域卷积操作的概念引入到图神经网络中, 首次提出了频域卷积神经网络模型 (spectral CNN). 自此, 在基于谱的图卷积网络方法得到了进一步改进和拓展[25−28]. 例如, ChebyNet[25]利用切比雪夫多项式的矩阵形式参数化核卷积, 极大地减少了 spectral CNN 参数量和计算复杂度, 从而使得谱方法变得实用起来. 但是, 谱方法在计算的时候通常需要同时处理整个图, 并且需要承担矩阵分解时的高时间复杂度, 难以并行或扩展到大图上. 因而, 基于空域的图卷积网络开始快速发展.

另一方面, 基于空间的方法直接在图结构上执行卷积操作, 将图卷积表示为从邻域聚合特征信息. GCN[17]作为空间方法的代表性工作, 对频域图卷积进行一阶近似来进一步简化, 使得图卷积的操作能够在空域进行, 极大地提升了图卷积模型的计算效率. 此外, 为了加快图网络的训练, GNN 还可以和采样策略相结合, 包括 SAGE[29]、FastGCN[30]、LADIES[31]等方法, 通过将计算限定在一个批量的节点而不是整个图中以实现高效计算的目的 (缓解训练时间和内存需求等问题). 随后, 为了使 GCN 更强大, 更多的基于空域的图卷积神经网络模型[32−36]被提出, 并在多种图数据相关的任务上取得了令人瞩目的成效. 鉴于该方法的自由度高、可计算性好、推理效率高等优点, 空间方法得到了广泛关注和发展. 另外, 有很多学者从不同角度 (如方法、应用等) 对图神经网络模型进行梳理和总结, 具体可以参考综述[37−47].

尽管图神经网络已经被证明是一种强大的非网格数据模型, 但是原始 GNN 依旧存在一定的局限性, 主要有两点: (1) 现有的 GNN 模型大多是半监督学习, 这就使得其性能会严重依赖于高质量标签数据; (2) 随着图数据规模的发展, 现有的图模型设计得越来越复杂, 这对图模型计算和图数据存储带来了一定挑战. 知识蒸馏在计算机视觉上的成功应用, 为上面两个挑战提供了一种可行的方案. 本文的第 1.2 节将简单回顾知识蒸馏在深度学习中的发展.

1.2 传统深度学习中知识蒸馏方法

知识蒸馏[12]最初被提出用于模型压缩, 不同于模型压缩中的剪枝和量化, 知识蒸馏 (knowledge distillation, KD) 采用教师-学生 (teacher-student, T-S) 模式预先训练一个大的教师模型来蒸馏得到一个轻量化的学生模型, 以此来增强学生模型的泛化能力, 达到更好的性能和精度. 通过蒸馏, 教师模型中的“知识”(软标签监督信息) 会转移到学生模型中. 通过这种方式, 学生模型可以减少时间和空间的复杂性, 还可以学习到在独热标签上学不到的软标签信息 (这些里面包含了类别间信息), 从而不会失去预测的质量. 通常, 根据知识迁移方式的不同, 知识蒸馏可以被分成 2 种技术路线.

第 1 种是目标蒸馏, 该方法与标签平滑密切相关[48], 利用教师模型的输出类别概率作为平滑标签来训练学生. 文献 [12] 是知识蒸馏的开山之作, 由 Hinton 等人在 2015 年提出, 首次提出将教师模型的 *Softmax* 层输出的类别的概率作为“Soft-target”迁移到学生模型中, 从而提高小网络性能. 为了学习学生网络中的反馈信息, DML[49]提出深度相互学习策略, 让一组学生网络同时训练, 通过真实标签的监督和同伴网络输出结果的学习经验实现相互学习共同进步. BAN[50]基于蒸馏的思想, 使用集成的思路训练学生模型, 使其网络结构和教师模型一样, 同时在计算机视觉和语言建模下游任务上明显优于其教师模型.

另一类知识蒸馏思路是特征蒸馏方法, 利用教师网络结构中的中间层特征表示所包含的语义信息作为知识迁移到学生模型中. FitNet[51]是最早采用这种方法的经典工作, 利用教师网络的输出和中间层的特征作为监督信息, 对文献 [12] 的知识蒸馏方法进行扩展, 实现深度模型网络压缩的问题. 基于特征蒸馏的方法现已成为主流, 包括注意力机制的使用[52], 特征空间的概率分布匹配[53,54]等. 在此之后, 基于特征蒸馏的方法还衍生出了一些基于关系蒸馏的新方法[55−61], 不过它们均旨在将 Teacher 中的特征知识迁移给 Student.

无论是哪一种蒸馏策略, 这些方法大多都是针对以网格数据为输入的卷积神经网络设计的. 幸运的是, 最近涌现出了在图和图神经网络上设计知识蒸馏的方法. 在接下来的内容中, 本文将简要梳理总结出在图上设计知识蒸馏算法的图蒸馏方法.

1.3 基于图的知识蒸馏发展和分类

随着知识蒸馏技术的发展, 仅蒸馏单个样本信息的蒸馏方法不再适用, 因为它们所提供的信息有限. 为了提取不同数据样本间的丰富相关性信息, 关系知识蒸馏方法[55−61]被提出, 这类方法通过隐式/显式构建样本间关系图, 充分挖掘教师网络中样本间的结构化特征知识. 随后, 图神经网络作为一种强大的非结构化建模工具, 可以直接在图关系数据上建模, 因此借助图神经网络进行蒸馏可以轻松实现图拓扑结构知识和样本间语义监督信息的提取和传递. 因此, 本文将基于深度神经网络的关系知识蒸馏方法和基于图神经网络的蒸馏方法统称为图知识蒸馏方法. 图知识蒸馏旨在将教师模型中直接/间接构造的样本关系语义信息蒸馏到学生模型中, 以获得更加通用的、丰富的、充分的知识.

虽然 GNN 这种强大架构在建模非结构化数据时有着很好的性能表现, 但 GNN 的卓越性能需要依靠高质量的标签数据和复杂的网络模型, 而标签数据获取困难且计算资源代价高昂. 所以, 面对图神经网络中存在的数据标签的稀疏性和模型计算的高复杂性问题, 如何在保证性能的情况下设计出更小、更快速的网络, 成为研究的重点. 基于这种思想, 在图数据上设计知识蒸馏算法的方法脱颖而出, 涌现出了各种各样的方法. 同时随着知识蒸馏在图分析任务上卓越的表现, 图知识蒸馏研究受到广泛的关注.

本文采用层次化的分类方法对图知识蒸馏方法[16,56−59,61−119]进行分类, 首先将图知识蒸馏方法分为面向深度神经网络的图知识蒸馏方法、面向图神经网络的图知识蒸馏方法和基于图知识的模型自蒸馏方法, 再进一步根据方法蒸馏位置的不同将其分为基于输出层、基于中间层和构造图的知识蒸馏方法, 具体分类以及代表方法如图 2 所

示. 需要注意的是, 在归纳总结基于图知识的模型自蒸馏方法时, 本文重点围绕图神经网络模型中的自蒸馏方法展开, 由于模型自蒸馏方法和 GNN 的结合近年来才被大家关注并研究, 因而方法相对于前两种比较少.

图 2　图知识蒸馏方法分类

2　基于 T-S 架构的图知识蒸馏方法

本节根据知识蒸馏算法设计的特点, 将图知识蒸馏方法分为面向深度神经网络的图知识蒸馏方法、面向图神经网络的图知识蒸馏方法和基于图知识的模型自蒸馏方法, 并对这 3 类方法分别进行介绍和分析. 其中, 表 1 汇总了全文中常见的符号及其表示的含义.

表 1　符号定义

符号	含义	符号	含义
X, Y	数据集, 标签	G, V, E	图, 节点集合, 边集合
n	样本个数	L	卷积层总数
W_t, W_s	教师和学生模型参数	h^l	第 l 层节点嵌入表示
z_t, z_s	教师和学生概率输出	v, u	目标节点, 邻居节点
p_t, p_s	教师和学生概率分布	x_t, x_s	教师和学生节点
τ	温度缩放系数	α	蒸馏损失超参数
L_{CE}	交叉熵损失	D_G	距离度量函数
L_{KD}	蒸馏损失	ψ_t	样本相似度函数
L_G	图蒸馏损失	S	节点相似度函数
L_{self}	图模型自蒸馏损失	G_t, G_s	节点间关系构造函数

2.1　面向深度神经网络的图知识蒸馏

首先, 为了简化表示, 本文将参数为 W_t 的表现良好的教师模型记做 T, 同样地, 将参数为 W_s 的学生网络模型记做 S . 卷积神经网络输入数据集记为 $X=\{x_1, x_2, \ldots, x_n\}$, 对应的标签记为 $Y=\{y_1, y_2, \ldots, y_n\}$, n 表示数据集中的样本个数. 由于深度神经网络可以看作是由多个非线性层叠加而成的映射函数, 则教师和学生的非归一化的概率输出分别记为 $z_t=\phi(X; W_t)$ 和 $z_s=\phi(X; W_s)$, 其中 ϕ 是映射函数. $p_t=Softmax(z_t)$ 和 $p_s=Softmax(z_s)$ 分别表示教师和学生的最终预测概率.

知识蒸馏最早由 Hinton 等人[12]提出, 目的是将隐藏在大型网络 (教师模型) 中的知识传递到轻量级小型网络 (学生模型) 中, 使得学生模型获得更好的性能. 知识蒸馏的基本思想是通过温度 τ 缩放 *Softmax* 输出层软化类概率分布, 得到软目标:

$$p^{\tau}=\frac{\exp\left(\frac{z_i}{\tau}\right)}{\sum_j \exp\left(\frac{z_j}{\tau}\right)} \tag{1}$$

其中, τ 表示温度缩放系数, 用于软化教师模型的输出, 当 τ 越大, 输出概率越平滑.

Hinton 等人[12]还发现了在训练过程中将软目标和真实标签一起指导学生模型, 会进一步提升学习效果, 具体是通过将这两部分的损失函数进行加权. 于是, 知识蒸馏的损失可以表示为:

$$L_{\mathrm{KD}}=L_{\mathrm{CE}}(p_s,y)+\alpha\tau^2\mathrm{KL}(p_s^{\tau},p_t^{\tau}) \tag{2}$$

其中, 第 1 部分 L_{CE} 就是传统的交叉熵损失, 即学生模型的预测输出 p_s 与真实标签 y 的交叉熵, 第 2 部分就是学生模型经过 τ 平滑后的预测输出 p_s^{τ} 与教师模型经过温度 τ 平滑后的预测输出 p_t^{τ} 的交叉熵, α 为调节两个损失函数比例的超参数, KL 为 Kullback-Leibler 散度.

然而, 传统的深度学习中的知识蒸馏方法 (如公式 (2) 所示) 大多都是针对单个样本进行学习. 为了进一步提升性能, 研究人员提出了特征蒸馏[51−54]和关系蒸馏[55−61]方法. 其中, 蒸馏效果最为显著的是关系知识蒸馏方法, 这类方法在本文中称为隐式构造图方法. 后续更是涌现出大量在中间卷积层或输出层中构建显式的样本关系图知识蒸馏方法, 尝试令学生模型去模拟教师模型中样本间的相似性而不再是模拟教师模型中单个样本的输出结果. 因此, 借助于构建的隐式/显式的样本关系辅助图, 学生模型可以充分挖掘教师网络中样本间的结构化特征信息, 实现从教师模型中提取通用的、丰富的、充分的知识来指导学生模型, 其概念图如图 3 所示.

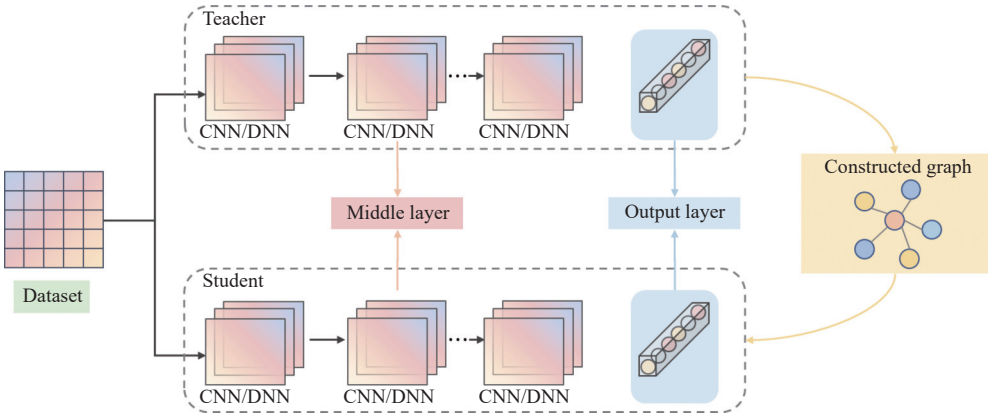


图 3 面向深度神经网络的图知识蒸馏方法框架

● 方法实现: (1) 首先, 分别针对 CNN/DNN 框架下的教师和学生模型得到的样本特征表示, 构造其各自的样本关系图 (如图 3 所示的 constructed graph 部分), 其中不同颜色的顶点表示不同的训练样本. (2) 其次, 使用相似性函数分别计算教师和学生网络样本间的相似性. (3) 最后, 使用距离度量函数最小化学生和教师的特征分布, 以保证学生模型可以学习多个输入样本在教师模型特征空间的相关性.

因此, 面向深度神经网络中的图知识蒸馏方法的最终的损失计算如公式 (3) 所示:

$$L_G=D_G\left(\psi_t\left(x_{s_i},x_{s_j}\right),\psi_s\left(x_{t_i},x_{t_j}\right)\right) \tag{3}$$

其中, x_{s_i} 和 x_{s_j} 表示学生网络中的两个输入样本 i,j, 同理 x_{t_i} 和 x_{t_j} 表示教师网络中的两个样本. $\psi_t(\cdot)$ 和 $\psi_s(\cdot)$ 分别表示学生和教师网络中样本间相似度函数 (如余弦相似度、Jaccard 相似度等), D_G 表示最小化学生和教师网络中构造图的距离度量函数, 可以是任意距离函数, 如欧氏距离、MSE、KL 等.

● 特点和优势: 隐式/显式构造图方法大多发生在中间卷积层上, 如图 3 所示, 借助于构建好的构造图, 学生模型可以将教师模型学习到的丰富样本间相关性知识直接提取到学习模型中, 而不再只是拟合教师模型中单个输入样本的输出类概率分布. 这样做的益处在于: 学生模型可以捕获到教师模型输入样本间的空间几何特征知识, 实现更准确地衡量样本特征之间的相似性, 进而提高学生模型的知识蒸馏学习效果. 在深度神经网络的知识蒸馏方法

中, 基于构造图进行知识蒸馏的方法[56−59,61−84]成为当下的研究热点, 相关工作将在第 3 节根据知识蒸馏位置的不同进行详细的分类和介绍. 然而, 如何正确地构造图来建模数据的结构知识仍然是一个具有挑战性的研究.

2.2 面向图神经网络的图知识蒸馏

本节首先对图神经网络的嵌入表示学习进行简化描述. 图作为一种常见的数据结构, 被用来广泛描述各种关系型数据. 其形式化表达如下, 图可以被表示为 $G=(V,E)$, 其中 V 是顶点集合, $|V|=N$ 表示图上共有 N 个节点, E 是边集合. 同时, 本文用 x 表示图 G 上的节点特征, 其中 $x\in R^{N\times D}, x_i\in R^D$ 表示第 i 个节点的特征, x_{ij} 表示第 i 个节点的第 j 个特征. 图表示学习主要是遵循消息传递范式[34]: 针对每一个节点 $v\in V$, 经过多层迭代聚合其邻域节点信息和节点本身的特征来更新节点的特征, l 次迭代后, 得到 v 的节点嵌入表示:

$$h_v^l=\sigma(h_v^{l-1}, AGG_{u\in N_v}\Phi(h_v^{l-1}, h_u^{l-1})) \tag{4}$$

其中, u 是节点 v 的邻居节点, N_v 是节点 v 的邻域节点集合; AGG 是聚合函数, 如 sum, mean 或 max; Φ 是消息函数, 如 MLPs (multilayer perceptrons); σ 是更新函数, 如 ReLU 激活函数.

尽管 GNN 现已成为当下图数据挖掘的研究热点, 并成功应用在医药、推荐、芯片设计等工业领域, 但 GNN 的性能严重依赖于大量高质量有标签数据和高度复杂的网络模型. 为了获得一个具有强泛化能力的 GNN 模型, 研究人员聚焦于在图数据上设计知识蒸馏算法, 将知识蒸馏与图神经网络相结合.

相比于深度神经网络中构建辅助图进行图知识蒸馏, 图神经网络作为一种强大的非结构化建模工具, 可以直接对图数据进行建模, 从而自然地挖掘教师模型中的图结构知识信息并传递到学生模型中. 面向图神经网络的图知识蒸馏方法类似于深度神经网络中的方法, 也是从图卷积中间层/输出层进行知识的提取. 为了进一步分析输入图数据结构节点间的关系, 在图神经网络中也会进一步借助构造图来学习教师模型图拓扑结构和节点关系信息, 以深入挖掘教师模型所学知识, 其概念图如图 4 所示.

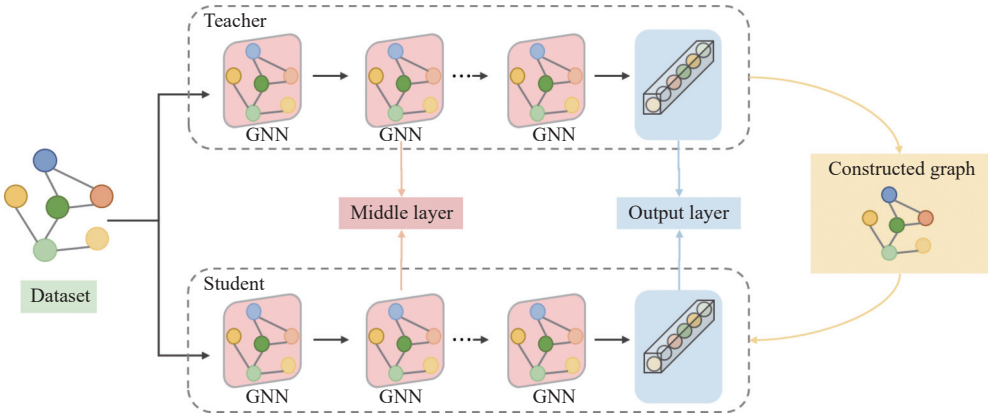


图 4 面向图神经网络的图知识蒸馏方法框架

● 同样地, 面向图神经网络的图知识蒸馏方法也主要分为以下 4 个步骤: (1) 首先, 分别基于 GNN 框架下教师和学生模型的中间特征表示, 构造其各自的节点间关系图 (如图 4 所示的 constructed graph 部分), 其中不同颜色的顶点表示不同的节点 (异构图中则表示不同类型的异构节点). (2) 其次, 使用相似性函数度量教师和学生网络内部拓扑结构节点间的相关性. (3) 然后, 使用距离度量函数计算教师和学生网络的各自内部节点表示之间的差异损失. (4) 最后, 将所有用于传递知识层的损失进行累加, 从而将图数据中的拓扑结构知识和样本关系知识迁移到学生模型中.

因此, 在图神经网络中, 图知识蒸馏最终的损失计算如下所示:

$$L_G=\sum_{l\in L}\sum_{(x,x')\in x^2} D_G\left(S\left(x_s^l, x_s'^l\right), S\left(x_t^l, x_t'^l\right)\right) \tag{5}$$

其中, x_s^l 和 $x_s'^l$ 表示 GNN 第 l 层中学生网络模型中的两个节点 x, x', 同理 x_t^l 和 $x_t'^l$ 表示 GNN 第 l 层教师网络模型中的两个节点. S 表示 GNN 卷积层/输出层中节点间相似度构造函数, D_G 表示最小化学生和教师中构造图的距离

度量函数, 如 Huber、MSE、KL、MAE 等.

● 特点和优势: 与面向深度神经网络的图知识蒸馏方法相比, 基于 GNN 的图蒸馏方法最大区别在于: GNN 是建模图数据的强大工具, 可以直接在中间层/输出层进行知识蒸馏, 就可以将图数据节点间的拓扑结构知识传授给学生模型. 后续为了进一步挖掘特征空间中局部样本节点间的关系, 不少工作尝试在中间图卷积层构造节点间的关系图来提取节点间的相关性知识传递给学生模型, 如 LSP[16]、HIRE[114]等. 知识蒸馏在图神经网络上的成功应用吸引了学术界和工业界的广泛关注, 涌现出了大量工作[16,85−114]. 本文将其统一归纳为面向图神经网络的图知识蒸馏方法, 相关工作将在第 4 节根据知识蒸馏位置的不同进行详细的划分和介绍. 然而, 在 GNN 中, 如何更充分地挖掘图拓扑结构和语义信息进行知识迁移仍然是一个具有挑战性的研究.

2.3 基于图知识的模型自蒸馏

自蒸馏是一种基于 T-S 架构的图知识蒸馏方法的特例, 它是指不借助额外教师模型下进行知识迁移的特殊蒸馏方式. 自蒸馏, 顾名思义, 即单个网络模型既是学生模型又是教师模型, 通常是将自身深层和浅层之间的信息进行传递来指导自身的学习, 而无需教师模型的辅助. 与两阶段式的 T-S 图蒸馏方法相比, 自蒸馏方法简单高效, 成为当前实际落地项目中的首选, 其概念图如图 5 所示 (本文只针对图神经网络模型中的自蒸馏方法进行总结, 不包括面向深度神经网络中的自蒸馏方法) .

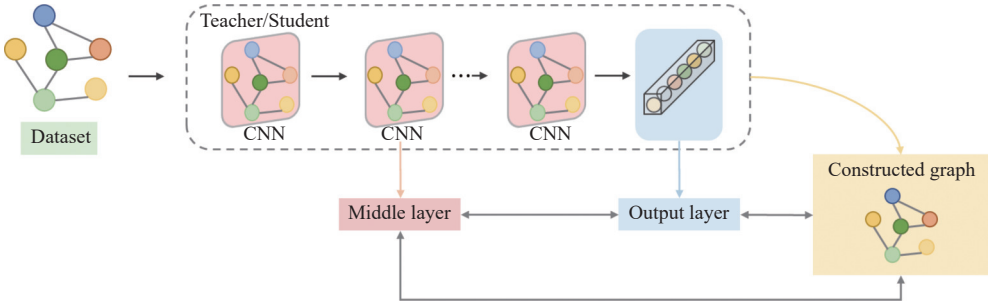


图 5 基于图知识的模型自蒸馏方法框架

● 基于图知识的模型自蒸馏方法实现步骤如下: (1) 首先, 基于 GNN 框架下的模型的中间层/输出层特征表示, 构造节点间关系图 (如图 5 所示的 constructed graph 部分), 其中不同颜色的顶点表示不同的节点 (异构图中则表示不同类型的异构节点). (2) 然后, 使用相似性函数度量 GNN 模型浅层和深层在特征空间中各自的内部拓扑结构节点表示间的相似性. (3) 最后, 使用距离度量函数计算浅层和深层网络的差异损失, 通过多次迭代计算可以学习到更加多样性的知识.

因此, 在图神经网络中, 基于图知识的模型自蒸馏方法整体损失如下所示:

$$L_{\text{self}} = \sum\nolimits_{l=1,\ldots,l-1} D_{\text{self}}\left(G_s^l(X), G_t^{l+1}(X)\right) \tag{6}$$

其中, X 表示整图节点表示, $G_s^l(X)$ 表示 GNN 浅层第 l 层中节点间关系知识构造函数, 同理 $G_t^{l+1}(X)$ 表示 GNN 深层第 $l+1$ 节点间关系知识构造函数, D_{self} 表示浅层和深层构造图的距离度量函数, 如 InfoCE、KL、MSE 等.

● 特点和优势: 与传统两阶段 T-S 蒸馏方式相比, 自蒸馏的学习模式可以极大地节省模型训练时间, 大幅提高训练的效率, 而且可以实现在无教师指导下达到模型性能的提升. 但是自蒸馏也存在一定的不足: (1) 缺乏了丰富的外部知识, 若可以显式地引入外部知识, 如与知识图谱结合使用, 或许对提升学生模型的性能有一定的帮助. (2) 传统两阶段的 T-S 蒸馏方式和自蒸馏方式孰优孰劣暂无结论, 目前缺少对它们在相同的实验环境和任务上的对比分析, 值得进一步研究. (3) 自蒸馏的理论分析还需进一步研究, 目前在 CNN/DNN 上的自蒸馏侧重于深层向浅层蒸馏效果, 而在 GNN 上则相反, 且缺乏理论支撑. 因此, 自蒸馏方法的通用性和灵活性有待进一步探讨.

3 面向深度神经网络的图知识蒸馏方法

知识蒸馏的核心在于知识的提取, 而知识存在于模型中的不同位置. 因此, 根据知识蒸馏的位置, 针对深度神

经网络的图蒸馏方法, 将其划分为输出层、中间层和构造图知识. 本节主要对这 3 类知识传递形式进行介绍, 下面介绍的相关图蒸馏方法都是以此为基础, 具体见表 2. 在具体图知识蒸馏方法描述时仅重点介绍各类图蒸馏算法最显著的知识蒸馏形式, 即若同时具有输出层、中间层知识的蒸馏方法默认划分为中间层知识蒸馏方法, 若同时具有这 3 类知识的蒸馏方法则默认划分为构造图知识蒸馏方法.

表 2　面向深度神经网络的图知识蒸馏方法汇总

方法	蒸馏位置			距离度量	任务	应用
	输出层	中间层	构造图			
IEP[67]	—	√	—	KL, L1	多任务学习	迁移学习, 图像分类
HKD[68]	—	√	—	InfoCE	知识蒸馏	图像分类, 知识迁移
CAG[75]	—	—	√	KL	图推理	视觉对话
DKWISL[62]	√	—	—	KL	自然语言处理	关系抽取
KTG[63]	√	—	—	KL	协同学习	图像识别
MHGD[69]	—	√	—	KL	多任务学习	图像识别
IRG[57]	√	√	—	Hit	知识蒸馏	图像识别
DGCN[64]	√	—	—	KL	协同过滤	商品推荐
GKD[76]	√	√	√	Frobenius	模型压缩	图像分类
SPG[65]	√	—	—	KL	自然语言处理	视频字幕
MorsE[77]	—	—	√	L2	元知识迁移	链接预测, 问答系统
GCLN[66]	√	—	—	L2	图像语义分割	视觉机器人自定位
DOD[70]	√	√	—	KL	目标检测	实例分割
BAF[78]	—	—	√	EMD	模型压缩	视频分类
LAD[79]	—	—	√	BELU	自然语言处理	机器翻译
GD[80]	—	—	√	余弦距离	多模态视频	动作检测, 动作分类
GCMT[81]	√	√	√	CE	无监督域适应	行人重识别
GraSSNet[82]	—	—	√	MSE	知识迁移	显著性预测
LSN[61]	√	√	√	KL, MSE	模型压缩	节点分类
IntRA-KD[83]	—	√	√	MSE	模型压缩	道路标记分割
RKD[56]	√	—	√	欧氏距离, Huber距离	知识蒸馏	图像分类, 少样本学习
CC[59]	√	√	√	KL, MSE	知识蒸馏	图像分类, 行人重识别
SPKD[58]	—	—	√	Frobenius	知识蒸馏	图像分类, 迁移学习
HKDIFM[71]	—	√	—	KL	知识蒸馏	图像分类
KDExplainer[72]	√	√	—	CE, KL	可解释性	图像分类
TDD[73]	√	√	—	CE, KL	可解释性	图像分类
DualDE[74]	—	√	—	JSD	知识蒸馏	节点分类, 链接预测
KCAN[84]	—	—	√	BPR	知识图谱	Top-K推荐, TR预测

● 相同点: 针对面向深度神经网络的图知识蒸馏方法中基于输出层、中间层和构造图这 3 类知识蒸馏形式, 每类图知识蒸馏算法的共性在于它们均是基于相同位置处知识的提取.

● 不同点: 针对每类图知识蒸馏算法, 它们的差异性体现在多个方面, 如具体方法实现、距离度量函数、下游任务及应用等方面上, 具体参见表 2. 如在基于输出层知识的图知识蒸馏这类方法中: DKWISL[62]采用 KL 距离度量方式将知识蒸馏应用在自然语言处理上的关系抽取中, KTG[63]使用 KL 度量教师和学生间的分布差异将知识蒸馏用于协同学习的图像识别应用上, 而 GCLN[66]则利用 L2 距离度量方式将知识蒸馏用于下游任务图像语义分割的视觉机器人自定位场景中. 在基于中间层知识的图知识蒸馏这类方法中: IEP[67]利用 KL 和 L1 相结合的距离度量方式将知识用于多任务学习上的迁移学习和图像分类中, HKD[68]使用 InfoCE 度量方法在图推理的视觉对话任务中引入知识蒸馏技术, 而 IRG[57]则采用 Hit 度量方式将知识蒸馏运用于图像识别场景上.

同样地, 基于构造图知识的图知识蒸馏方法也是类似的: CAG[75]采用构造图知识蒸馏技术提升学生模型在下游图推理任务上的视觉对话表现, GKD[76]利用 Frobenius 最小化教师和学生间的分布差异并对学生模型进行压缩, MorsE[77]则使用 L2 度量方法进行元知识的迁移以提升学生模型在链接预测和问答系统上的表现.

3.1 基于输出层知识

基于输出层知识是网络模型最后一层预测输出所蕴含的标签监督信息, 是目前深度神经网络中最流行的知识蒸馏形式, 自 KD[12]方法被提出以来, 这类方法便受到学者的广泛关注, 随后衍生出了大量优秀工作, 本节重点关注 CNN/DNN 上输出层借助于辅助图的图蒸馏方法, 即考虑样本间关系的图知识蒸馏方法.

最早的输出层图知识蒸馏可以追溯到 Minami 等人[63]提出基于图来控制知识转移的方法 KTG, 通过知识转移统一视图, 来表示不同的知识转移模式, 同时还引入 4 种类型的门函数控制网络训练时的反向传播, 来探索不同的知识转移组合形式. 同年, Wang 等人[64]借助于 GNN 模型, 将 GCN 模型中的排名信息蒸馏到二进制模型中, 以充分挖掘商品数据中用户和商品间的丰富连接信息, 成功实现在线推荐模型的性能提升和隐式反馈推荐的加速. 同时, Zhang 等人[62]也提出将输出层软标签和 GCN 模型结合使用, 成功将输出层知识应用在自然语言处理领域中. 具体地, DKWISL 首先从整个语料库获得类型限制的 soft rules, 而后教师模型将设计的 soft rules 与 GCN 相结合并针对每个实例得到最终的软标签, 其模型架构见图 6.

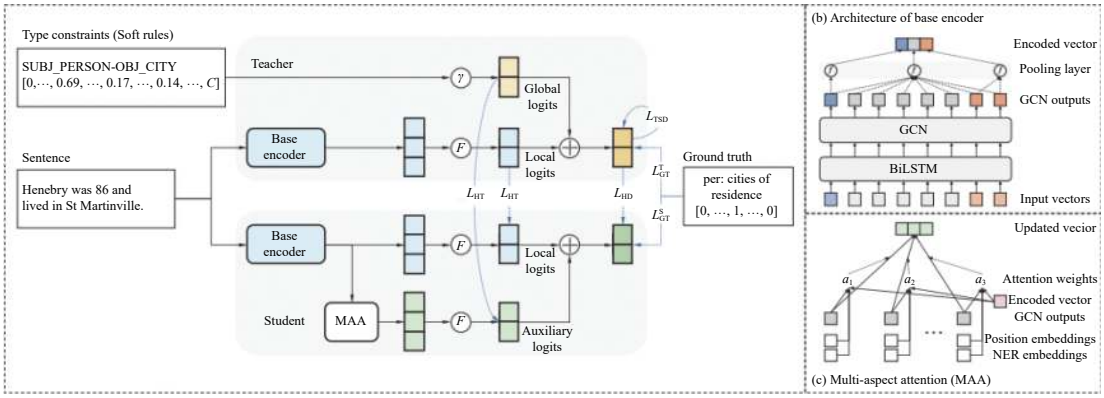


图 6 DKWISL[62]的输出层知识蒸馏

此外, 还有一些相关工作被提出. Pan 等人[65]认为以往视频描述模型没有清晰地建模对象间的相互作用, 提出一个新颖时空图网络明确利用时空对象交互, 同时引入一种具有对象感知的知识蒸馏机制 SPG, 利用局部对象信息对全局场景特征进行正则化, 解决了时空图模型中存在的噪声特征学习问题. Koji 等人[66]将输出层知识蒸馏应用于机器人自定位应用中, 即通过设计一种基于秩匹配的教师-学生知识转移方案 GCLN, 将现有的教师自定位模型的倒数秩向量输出作为暗知识转移到学生模型中.

3.2 基于中间层知识

考虑到输出层知识蒸馏方式的单一性, 为了进一步挖掘教师网络中蕴含的丰富知识, 很多研究学者开始研究如何将中间卷积层中的特征知识也转移到学生网络中, 以获得更充分的特征表示. FitNets[51]是最先使用中间层特征蒸馏的方法, 旨在利用教师模型特征提取器的中间层输出作为 hints, 对更深更窄的学生模型进行知识蒸馏. 不同于简单的中间层特征蒸馏. 本节重点介绍基于中间层特征关系知识, 汇总在深度神经网络上进行蒸馏的方法. 在这类方法中, 最具有代表的工作是 Liu 等人[57]在 2019 年 CVPR 上提出的 IRG 方法 (见图 7).

从图 7 得知, 不同于只考虑实例特征知识, IRG 还额外引入实例关系、特征空间变化这两种知识. 具体地, 该模型为了建模网络中间层的抽象知识, 通过将实例特征和实例关系分别作为顶点和边来构建一个实例关系图 IRG, 并基于此提出 IRG 变换的跨层特征空间的图蒸馏方法, 实验验证了该方法有效地捕获整个网络上的知识, 同时对不同的网络结构具有较强的鲁棒性.

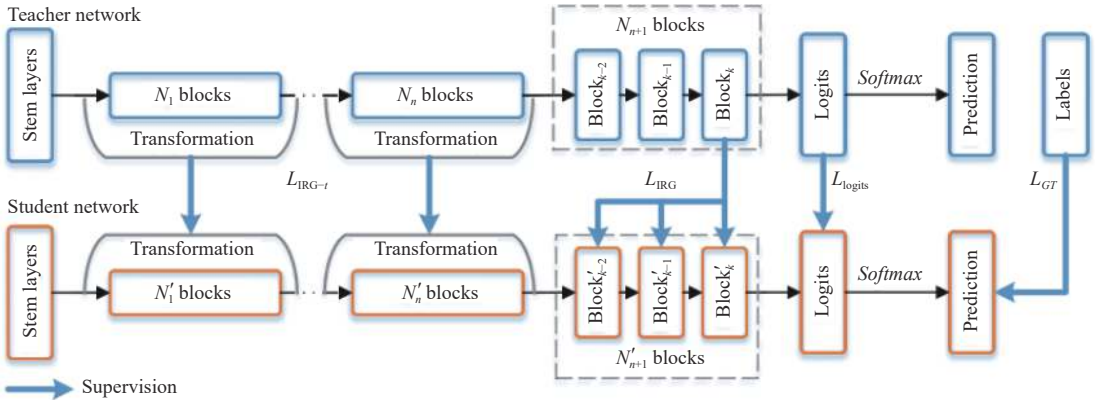


图 7 IRG[57]的中间层知识蒸馏

IRG 图蒸馏方法的成功, 催生了大量相关工作. 比如, Lee 等人[69]提出一种利用多头注意机制将教师嵌入过程中的知识提取为图, 并通过多任务学习使学生具有关系归纳偏置的能力的 MHGD 图知识蒸馏方法. Passalis 等人[71]认为现有的 KD 方法通常忽略了神经网络在训练过程中经历的不同学习阶段, 提出在关键学习阶段令学生模型模拟教师模型的信息流, 以确保网络各层次之间形成必要的连接, 旨在实现知识的有效传递. Lee 等人[67]认为良好的知识应该能够解释嵌入过程, 于是提出一种基于主成分分析的可解释嵌入过程 (IEP) 知识生成方法, 利用 MPNN[34]提取该知识, 并通过可视化证实了 IEP 对嵌入过程知识的可解释性. Zhou 等人[68]认为当前个体知识和关系知识之间的关系被忽略, 提出基于实例间的属性图来从教师网络中提取整体知识, 实现个体知识和关系知识的融合, 同时保留了两种知识之间的相关性, 使得学生模型在训练时获得充足的知识. Xue 等人[72]提出 KDExplainer 来阐明在 KD 过程中软目标的工作机制, 发现 KD 可以隐式调节子任务间的知识冲突, 比标签平滑更有效, 同时基于这些观察结果, 还提出了一个便携式紧凑模块 VAM, 进一步改善基础 KD 的结果. Song 等人[73]为了弄清楚预训练的教师模型背后解决问题的过程, 提出了一种新的树状决策蒸馏 (TDD), 通过 layer-wise 方式剖析教师决策过程, 并将相同的决策约束强加于学生模型, 促使学生掌握相同的问题解决方案, 最终, TDD 成功探索了教师的决策过程, 发现教师的决策过程以从粗 (浅层获得粗粒度判别) 到细 (深层获得细粒度判别) 的方式进行. Zhu 等人[74]在蒸馏过程中考虑了教师和学生之间的双重影响, 提出了一种名为 DualDE 的新型知识图谱表示蒸馏方法, 通过将教师和学生的三元组输出分数和中间层嵌入结构知识均蒸馏给彼此, 并引入软标签评估机制来评估教师/学生提供的软标签质量, 以实现学生和教师的双重优化. Chen 等人[70]基于每个感兴趣区 (RoI) 实例之间的关系设计了一个结构化实例图, 同时利用实例特征和特征间的相似性, 以结构化的方式传递知识, 保证学生模型可以捕获全局拓扑结构知识和软标签知识.

基于中间层知识的蒸馏方法不仅可以蒸馏单个样本的知识, 还可以蒸馏特征空间样本关系知识到学生网络中, 成为当前深度神经网络中图知识蒸馏的重要方法之一.

3.3 基于构造图知识

为了更好地建模教师网络中样本间关系监督信息, 构造图知识蒸馏方法被提出. 基于构造图的知识蒸馏方法是中间层图知识方法的扩展, 旨在通过构造显式的辅助图结构模块来深入挖掘教师网络中样本特征间的高阶关联知识. RKD[56]是该类方法的代表工作, 由 Park 等人在 2019 年提出, 通过实验证明了提取样本间的关系结构信息优于提取单个样本的特征信息, 如图 8 所示.

RKD 这种通过以模型输出的结构信息进行蒸馏的方式被提出后, 基于构造图的知识蒸馏方法成为面向深度神经网络的图知识蒸馏研究热点. 与之相关的基于构造图知识工作应运而生. 例如, Zhang 等人[78]认为当前视频表示学习局限在一个学习任务中且计算代价高昂, 提出利用 logits 和中间特征构造两个辅助图, 将多个自监督教师的知识传递给学生, 实现模型压缩的目标. Chen 等人[61]从特征嵌入的新视角看待师生蒸馏范式, 通过引入局部位置保留损失 LSN, 借助图来维持教师网络中高维空间中样本之间的关系知识, 鼓励学生网络生成低维特征, 以保

留教师网络中对应的高维特征的实例间的关系信息. Peng 等人[59]认为除了实例一致性之外, 实例之间的相关性也是增强学生表现的有价值知识. 于是提出相关一致性知识蒸馏框架 CC, 该框架在传递知识时不仅考虑了输出层知识和中间特征知识, 还使用了基于图的知识 (实例间的关联信息). Tung 等人[58]观察到相似语义的输入往往会引发相似的激活模式, 因此引入了保持相似性知识蒸馏 SPKD 指导学生网络的训练, 使得学生不需要模仿教师的表征空间, 只需在其自身表征空间中保留成对的相似性. Lassance 等人[76]扩展了 RKD 方法, 借助图捕获潜在空间的几何信息, 从而图将知识从教师架构转移到学生网络中.

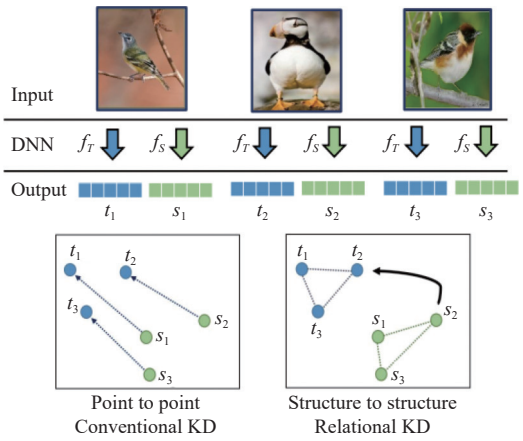


图 8 RKD[56]的核心思想

另外, 除了将构造图知识蒸馏用于模型压缩和模型增强外, 很多学者探究 KD 在其他领域上的应用. He 等人[79]引入语言图的概念, 并进一步提出了图蒸馏算法 LAD 来提高机器翻译的准确性. Luo 等人[80]提出跨模态动态蒸馏方法 GD, 将图蒸馏和域迁移技术相结合, 先在源域上训练, 然后固定基础模型在目标域上进行蒸馏, 实现多模态之间的知识迁移. Liu 等人[81]针对无监督域自适应行人重识别任务, 通过在教师和学生网络之间构建图一致性约束, 提出了一种基于图一致性的均值教学 (GCMT) 方法, 有效地优化了包含更多样本相似关系的学生网络的蒸馏训练过程. Zhang 等人[82]一种新的图语义显著性网络 (GraSSNet), 借助于构造的辅助图, 对从外部知识中学习到的语义关系进行编码, 用于实现显著性预测. Tu 等人[84]研究了如何引入外部知识图谱结构信息到推荐网络中, 通过局部条件注意力在采样子图上传播个性化信息来蒸馏知识图谱. Chen 等人[77]针对知识图谱任务 (KGs), 提出了一种新的任务元知识转移方法 MorsE, 将元知识从构造的源 KGs 转移到新的目标 KGs, 在链接预测和问答系统上均获得了优异的表现. Guo 等人[75]提出了一个用于视觉对话的细粒度上下文感知图 (CAG) 蒸馏方案, 如图 9 所示, 旨在通过发现部分相关的情境并构建适当的图结构, 并从中获得广义意识能力, 来解决视觉对话任务中存在的噪声问题. Hou 等人[83]将构造图知识蒸馏成功应用在道路标记分割场景中, 如图 10 所示, 基于构建的 affinity graph 特征相似图表征不同类型的道路标记, 即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节点, 然后根据特征分布的相似性建立节点间的成对关系, 将结构知识提取到学生网络中.

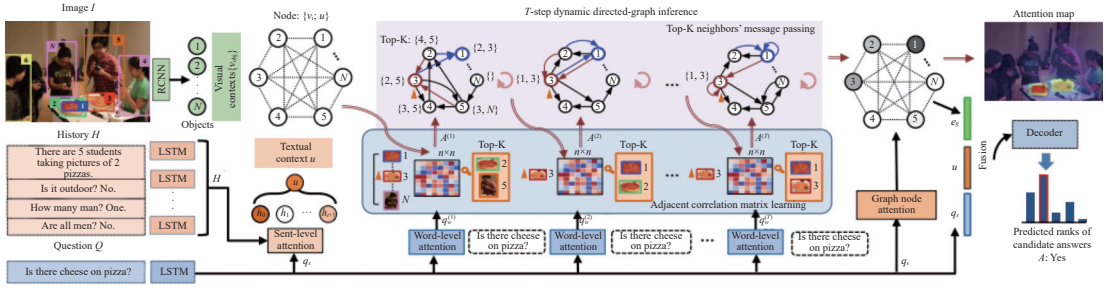


图 9 CAG[75]的构造图知识蒸馏用于视觉对话

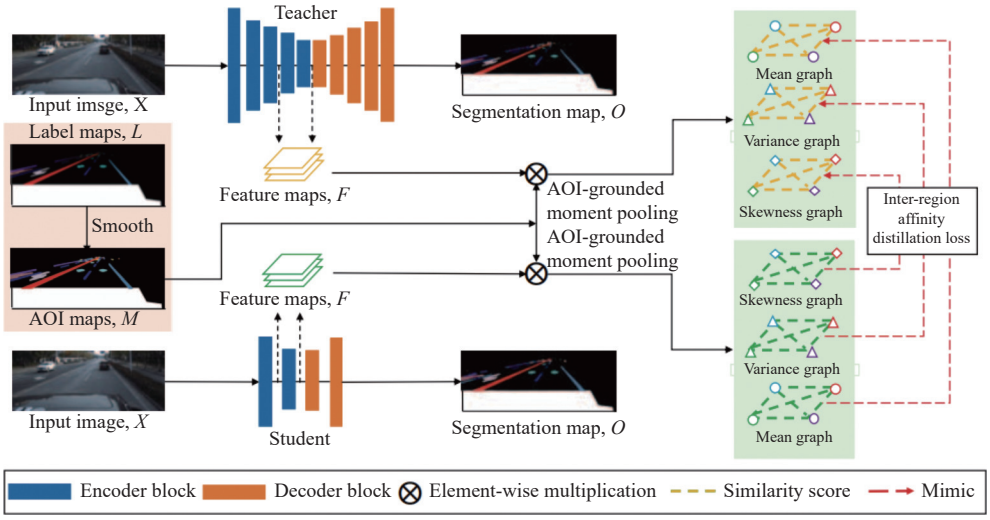


图 10 IntRA-KD[83]的构造图知识蒸馏用于道路标记分割

尽管面向深度神经网络的图知识蒸馏方法展现出了巨大的潜力, 并成功应用在多种下游任务中. 但是这类方法聚焦于 CNN/DNN, 无法直接应用具有图结构形式的数据上. 近年来, 大量研究人员尝试将 KD 应用在 GNN 上, 并取得了令人印象深刻的成果, 这部分内容将在第 4 节进行具体阐述.

4 面向图神经网络的图知识蒸馏方法

类似地, 针对图神经网络中的知识蒸馏方法, 本节仍依照知识蒸馏的位置, 将其分类为基于输出层知识、基于中间层知识和基于构造图知识. 本节对面向图神经网络的图知识蒸馏方法进行细致划分, 具体如表 3 所示. 同样地, 这里仅重点介绍各类图蒸馏算法最显著的知识蒸馏形式, 即若同时具有输出层、中间层知识的蒸馏方法默认划分为中间层知识蒸馏方法, 若同时具有这 3 类知识的蒸馏方法则默认划分为构造图知识蒸馏方法.

● 相同点: 针对面向图神经网络的图知识蒸馏方法中基于输出层、中间层、和构造图这 3 类知识蒸馏形式, 每类图知识蒸馏算法的共性在于它们均是基于相同位置处知识的提取.

● 不同点: 针对每类图知识蒸馏算法, 它们的差异性体现在多个方面, 如具体方法实现、距离度量函数、下游任务及应用等方面上, 具体参见表 3. 如在图知识蒸馏中用 KL 散度的, 基于输出层有 GFKD[85], GLNN[88], Distill2Vec[89]和 MT-GCN[90]等; 基于中间层有 MustaD[97], OAD[102], BGNN[104]和 HSKDM[106]; 基于构造图有 CPF[110], LSP[16], MetaHG[112]和 HIRE[114]. 同样地, 用 MSE 度量方式的, 基于输出层有 RDD[86]和 SCR[93]; 基于中间层有 LWC-KD[96], AGNN[99], Cold Brew[100]和 EGSC[105]; 基于构造图有 HIRE[114]. 此外, 还有一些使用其他距离度量函数来度量教师和学生模型之间分布差异的, 具体统计见表 3.

表 3 面向图神经网络的图知识蒸馏方法汇总

方法	蒸馏位置			距离度量	任务	应用
	输出层	中间层	构造图			
GFKD[85]	√	—	—	KL	无数据图蒸馏	零样本学习
LWC-KD[96]	—	√	—	MSE, L2	增量学习	推荐系统
MustaD[97]	√	√	—	KL	模型压缩	节点分类
RDD[86]	√	—	—	MSE	半监督学习	节点分类
EGAD[98]	—	√	—	RMSE, MAE	半监督学习	实时视频流事件预测
GRL[107]	—	—	√	MAE	多任务学习	图级预测
GFL[108]	—	—	√	Frobenius	小样本学习	节点分类

表 3 面向图神经网络的图知识蒸馏方法汇总 (续)

方法	蒸馏位置			距离度量	任务	应用
	输出层	中间层	构造图			
HGKT[109]	—	—	√	Wasserstein	零样本学习	节点分类
AGNN[99]	√	√	—	MSE	模型压缩	节点分类, 点云分类
CPF[110]	√	√	√	L2, KL	知识蒸馏	节点分类
LSP[16]	—	√	√	KL	模型压缩	节点分类, 点云分类
GKD[87]	√	—	—	CE	图推理	疾病诊断预测
scGCN[111]	—	—	√	CE	单细胞组学	细胞识别, 跨物种分类
MetaHG[112]	—	—	√	KL	非法毒品贩子检测	分类
Cold Brew[100]	√	√	—	MSE, CE	冷启动	推荐系统
PGD[101]	—	√	—	MSE	冷启动	推荐系统
GLNN[88]	√	—	—	KL	离线知识蒸馏	节点分类
Distill2Vec[89]	√	—	—	KL	模型压缩	链接预测
MT-GCN[90]	√	—	—	KL	半监督学习	节点分类
TinyGNN[91]	√	—	—	CE	模型压缩	节点分类
GLocalKD[92]	√	—	—	KL	异常检测	异常检测
OAD[102]	√	√	—	CE, KL	在线对抗蒸馏	节点分类
SCR[93]	√	—	—	MSE	模型训练	节点分类
ROD[94]	√	—	—	KL	模型压缩	节点分类、聚类, 链接预测
EGNN[95]	√	—	—	KL	模型可解释性	节点分类
CKD[103]	—	√	—	JSD	知识蒸馏	节点分类, 链接预测
G-CRD[113]	√	√	√	InfoCE	模型压缩	分类, 相似性度量
BGNN[104]	—	√	—	KL	模型压缩	图像分类
EGSC[105]	—	√	—	MSE, Huber	模型压缩	异常检测, 图相似性计算
HSKDM[106]	—	√	—	KL, Triplet	知识蒸馏	节点分类
HIRE[114]	√	√	√	KL, MSE	知识蒸馏	节点分类、聚类、可视化

4.1 基于输出层知识

受深度学习中图蒸馏技术的启发, 图神经网络相关研究人员将 KD 和 GNN 技术相结合, 已经成功应用在图数据挖掘的各种应用场景中, 包括推荐学习、异常检测、和细胞识别等. 本节重点关注 GNN 中基于输出层知识蒸馏方法.

KD 在 GNN 上的应用近年来才开始受到大家关注, 输出层蒸馏工作最早可以追溯到 2020 年 TinyGNN[91]工作的提出, 其蒸馏训练过程可见图 11. TinyGNN 是为了解决浅层 GNN 和深层 GNN 之间的邻居信息差距. 具体地, Yan 等人提出了对等感知模块 (PAM) 和邻居蒸馏策略 (NDS), 分别显式和隐式建模局部节点信息结构. 通过这种方式, TinyGNN 可以有效地表征局部结构, 学习到更好的节点表示, 从而达到和深层 GNNs 相同甚至更好的表现.

随后, 研究人员提出大量基于输出层知识蒸馏的相关工作. 譬如, Zhang 等人[86]认为教师模型预测不可靠将导致额外的计算开销和高偏差, 设计了可靠数据蒸馏 (RDD) 方法, 通过定义图中节点的可靠性和边的可靠性以更好地利用高质量的数据, 从而优化传统 KD 并增强模型表征能力. Antaris 等人[89]设计了一个基于 Kullback-Leibler 散度的蒸馏损失函数, 将获取的知识从基于离线数据训练的教师模型转移到基于在线数据训练的小型学生模型中, 同时还采用了一种自注意机制来捕捉学习到的节点嵌入中的图演化. Deng 等人[85]考虑到因数据隐私等问题导致的数据不可用问题, 设计一种从无图数据的 GNN 中进行知识蒸馏的 GFKD 框架. 具体地, GFKD 首先学习教师 GNN 模型中知识更容易集中的 fake graphs, 然后利用这些 fake graphs 将知识迁移给 GNN 学生模型. 为了实现这一目标, GFKD 提出了一种结构学习策略, 对多元伯努利分布的图拓扑进行建模, 然后引入梯度估计对其进行优

化. Ghorbani 等人[87]使用标签传播算法将所有与图相关的知识注入到教师模型产生的伪标签中, 然后使用该伪标签训练学生网络. Zhan 等人[90]提出了一种简单而有效的基于图半监督学习的蒸馏方法 MT-GCN, 将高置信度预测作为伪标签来扩展标签集, 以便选择更多的样本来更新 GCN 模型. Zhang 等人[93]通过知识蒸馏来研究一致性正则化在半监督图神经网络训练中的作用, 提出可以通过计算学生和教师模型之间的一致性损失来指导模型的训练. Zhang 等人[94]为了解决 GNN 边缘稀疏性和标签稀疏性问题, 首次提出在线蒸馏方法 ROD, 通过整合多尺度感知图知识来动态训练强大的教师/学生模型. Ma 等人[92]提出了一种新颖的深度异常检测方法 GLocalKD, 该方法通过对图和节点的联合随机蒸馏来学习图数据丰富的全局和局部的图正则化模式信息. Li 等人[95]基于知识蒸馏思想, 提出一种用于图表示的可解释浅层图神经网络模型 EGNN. 具体地, EGNN 结合知识蒸馏, 同时通过设计透明、可解释的邻居选择策略, 来解释图神经网络模型的聚合操作.

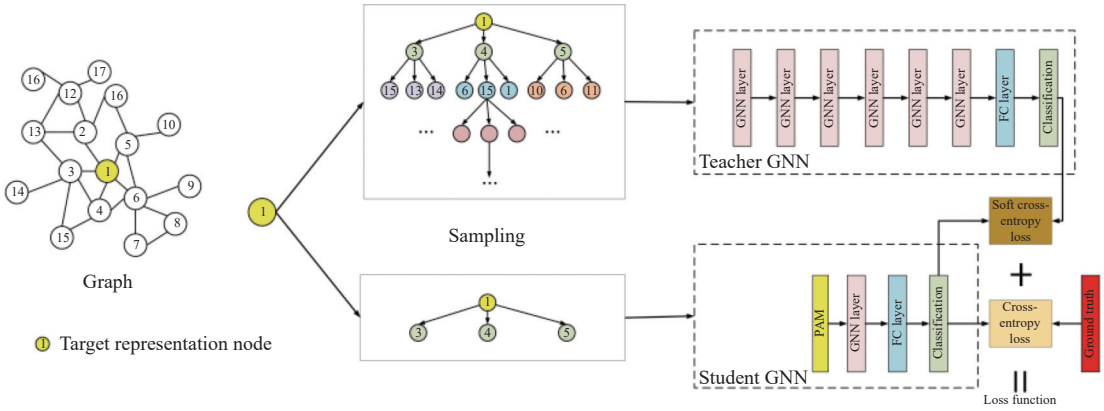


图 11 TinyGNN[91]的输出层知识蒸馏

不同于上面的师生模型框架一致的知识蒸馏方案, Zhang 等人[88]提出一种新的蒸馏框架 GLNN, 通过将教师 GNN 中的 logits 知识提取到学生 MLP 模型中, 大大减少了节点分类的推理时间. 图 12 展示了 GLNN 的模型框架图.

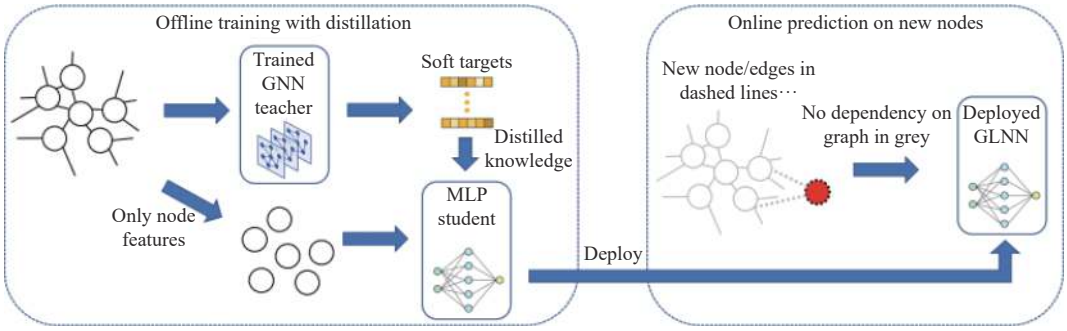


图 12 GLNN[88]的输出层知识蒸馏

具体来说, GLNN 是一个从 GNN 到 MLP 的 KD 蒸馏范式, 由此得到的 GLNN 是通过 KD 优化的 MLP, 其在训练中具备图上下文感知能力, 且在推理过程中没有图的依赖性, 推理速度与 MLP 一样快. 因此, GLNN 模型既有 MLP 的低延迟和无图依赖性的优点, 又可以达到和 GNN 相同的表现, 从而解决 GNN 延迟性高的问题.

虽然基于输出层知识在 GNN 上展现出了巨大的优势, 但是这类方法只是套用在了文献 [12] 提出的 KD 框架上, 仅利用了输出层类别概率分布这一种监督信息, 未能充分挖掘教师 GNN 中的知识.

4.2 基于中间层知识

鉴于 GNN 中输出层知识对教师模型信息提取的有限性, 中间层知识的引入可以进一步丰富知识的表示形式

且能够有效提高学生模型下游任务的性能. 本节对面向图神经网络中的中间层知识蒸馏方法进行汇总, 这类方法根据下游任务, 主要分为两大类.

一类是用于常见的图挖掘任务中的节点分类等任务中. 例如, Jing 等人[99]提出多对一的师生蒸馏框架 AGNN (见图 13), 即用多个教师网络来共同训练一个学生网络, 使学生能够从具有不同特征维度的教师那里学习. Kim 等人[97]整合教师网络中间图卷积层的聚合信息和输出层的软标签知识, 提取到学生网络中. Wang 等人[102]为 GNN 设计出首个在线对抗蒸馏方法 OAD, 核心思想是在知识蒸馏训练过程中引入生成对抗学习和循环学习技术, 以有效捕捉图神经网络的结构变化. Wang 等人[103]首次尝试用协同知识精馏方法对异构信息网络中元路径嵌入之间的相关性进行建模, 可以有效地在最终嵌入过程中保持全局相似度和蒸馏局部知识. Huang 等人[106]为了提高 GNN 在类别不平衡图数据中的分类性能, 提出了一种基于硬样本的知识蒸馏方法 HSKDM, 通过联合训练多个 GNN 模型, 将模型的中间层和输出层知识一起提取到学生网络中, 最后显著提高了节点分类性能.

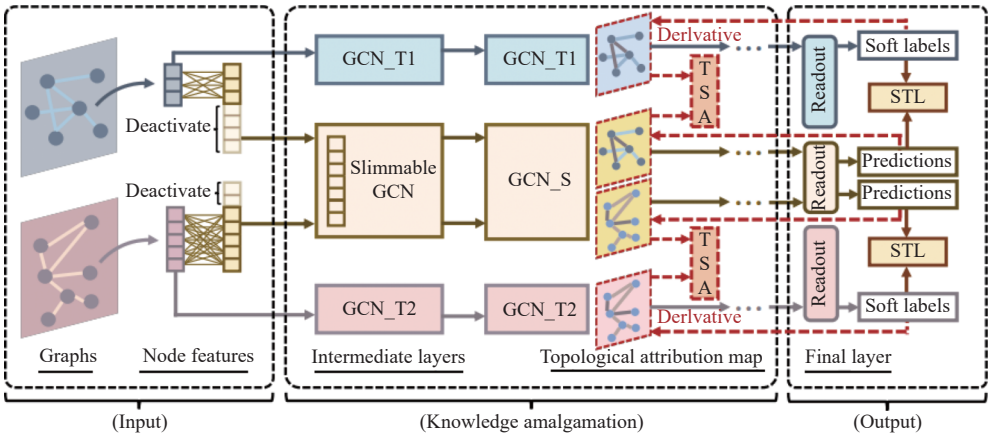


图 13 AGNN[99]的中间层知识蒸馏

另一类是将中间层知识图蒸馏方法用于其他图分析任务, 如推荐系统[96,100,101]等任务中. 如图 14 所示, Zheng 等人[100]采用知识蒸馏技术, 通过利用图结构将教师节点嵌入到低维流形上, 并要求学生学习从节点特征到该流形的映射, 而将 GNN 推广到推荐系统中的冷启动研究问题中. 同年, Wang 等人[96]将对比学习和蒸馏学习相结合, 提出逐层对比蒸馏框架 LWC-KD 用于增量学习的推荐场景中. Wang 等人[101]利用图学习和知识蒸馏在特权信息建模中的优势, 提出了一种新的特权图蒸馏模型 (PGD), 以提高 GNN 模型在冷启动问题上的性能表现. 此外, GNN 的中间层知识蒸馏还应用于其他任务. 如 Antaris 等人[98]首次研究并提出动态图表示学习的知识蒸馏方法 EGAD, 在连续动态图卷积网络之间引入带权重的自注意机制来捕捉实时视频流事件图的演化, 并准确地学习潜在节点表示. Bahri 等人[104]介绍了一种基于 XNOR-Net++和知识蒸馏的二值化图神经网络 BGNN, 通过研究各种策略和设计决策来探究对二值化图神经网络图像分类性能的影响. Qin 等人[105]针对图相似度学习缓慢的问题, 设计了一种新颖的基于共同注意的多级 GNN 特征融合模型, 并采用知识蒸馏的方法从该模型中提取知识蒸馏到学生模型.

GNN 上的中间层图蒸馏方法的在各种图分析任务中的成功应用, 表明了中间层知识对图蒸馏技术的重要性.

4.3 基于构造图知识

为了进一步丰富并提供更通用的知识, 在 GNN 中也会进一步借助构造图[16,107−114]来学习教师模型图拓扑结构和节点关系信息, 以深入挖掘教师模型中蕴含的知识. 在这类方法中, LSP[16]是首个专门为同构图神经网络设计的知识蒸馏框架, 图 15 展示了 LSP 的蒸馏过程.

如图 15 所示, Yang 等人[16]在 GNN 中设计了一个局部结构保持模块 LSP 来捕获图拓扑信息, 将来自教师和学生的局部结构信息建模为分布信息, 通过最小化分布之间的距离实现从教师到学生模型的知识迁移. 受到 LSP 方

法的启发, CPF 蒸馏方法[110]被提出, 该方法将学生模型设计为参数化标签传播和特征转换模块的可训练组合, 使得学生可以从教师模型中基于结构和特征的先验知识中获益, 从而比教师模型具有更好的预测表现. Joshi 等人[113]在 LSP 的基础上, 设计新型图对比学习表示蒸馏框架 G-CRD, 该框架基于对比学习的思想对齐教师与学生的节点嵌入表示来隐式保留全局拓扑结构. Song 等人[111]提出一种单细胞图卷积网络模型 scGCN, 并结合知识蒸馏技术, 实现跨越不同数据集之间的有效知识转移. 具体地, scGCN 通过预处理, 将稀疏的原始数据集转变为包含跨数据相关信息的映射图, 这使得在参考数据集和未知数据集之间共享信息, 识别标签间相互关系, 并迁移到未知数据集上成为可能. Qian 等人[112]提出了一个元学习蒸馏框架 MetaHG, 旨在通过联合建模社交媒体上结构化关系和非结构化内容信息为异构图, 并引入元学习将图结构知识从训练任务中转移并有效泛化到下游测试非法毒品交易任务中, 以解决标签稀疏问题. Ma 等人[107]提出了一种新的多任务知识蒸馏方法 (GRL) 用于图级表示学习, 该方法通过多任务学习, 将基于网络理论的图度量作为辅助任务来学习更好的图表示. Yao 等人[108]提出了图少样本学习模型 GFL, 为了更好地捕捉全局信息, 借助于构造的辅助图结构学习一个可迁移的度量空间, 同时蒸馏局部节点级和全局图级关系结构知识到模型中. Wang 等人[109]提出了一种基于异构图的知识转移方法 (HGKT), 借助于构建的结构化异构图, 同时捕获类间和类内关系, 通过转移不可见类的邻接类的知识, 来计算不可见类的节点表示.

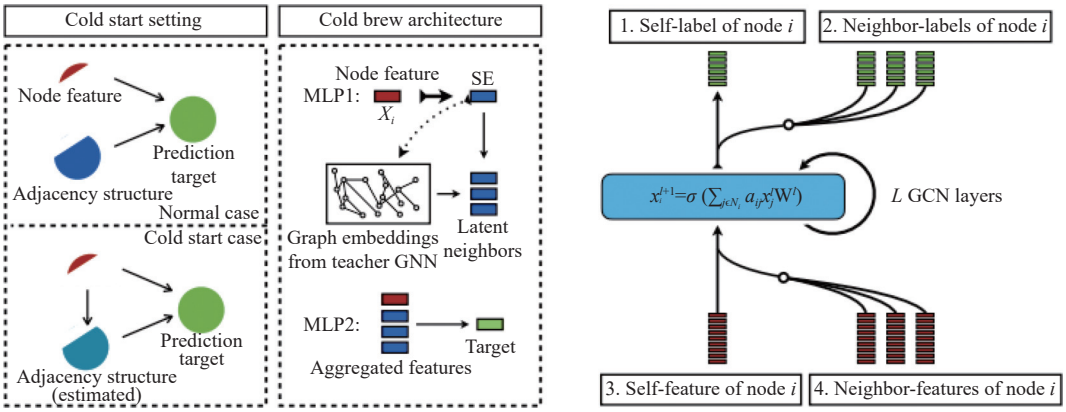


图 14 Cold Brew[100]的中间层知识蒸馏

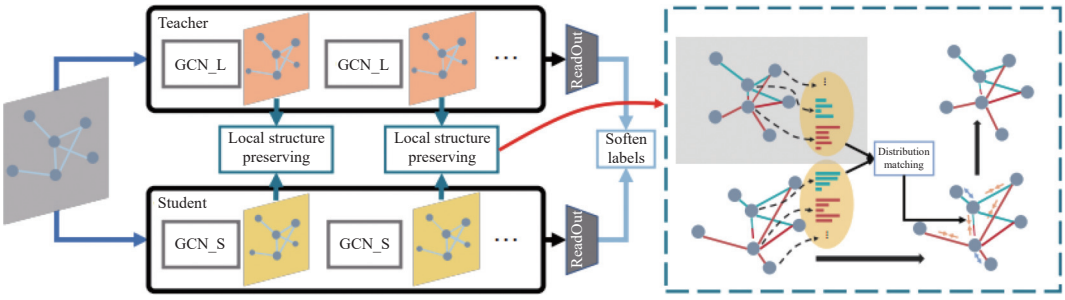


图 15 LSP[16]的构造图知识蒸馏

此外, 除了针对同构图的知识蒸馏模型, Liu 等人[114]还提出了 HIRE, 专门为 HGNN 设计的高阶关系知识蒸馏框架, 该方法成为一种实用且通用的训练方法, 适用于任意的异构图神经网络, 不仅提升了异构学生模型的性能和泛化能力, 而且保证了对异构图神经网络的节点级和关系级知识提取. 图 16 展示了 HIRE 基于构造图的师生蒸馏过程.

如图 16 所示, HIRE 方法主要由两部分组成, 包括节点级知识蒸馏 (node-level knowledge distillation, NKD) 和关系级知识蒸馏 (relation-level knowledge distillatio, RKD). 其中, NKD 用以对预训练异构教师模型的单个节点语义进行编码; RKD 用来对预训练异构教师模型的不同类型节点之间的语义关系进行建模. 最后, HIRE 通过整合节点级知识蒸馏和系级知识蒸馏, 同时考虑单个节点软标签和不同节点类型之间的相关性知识.

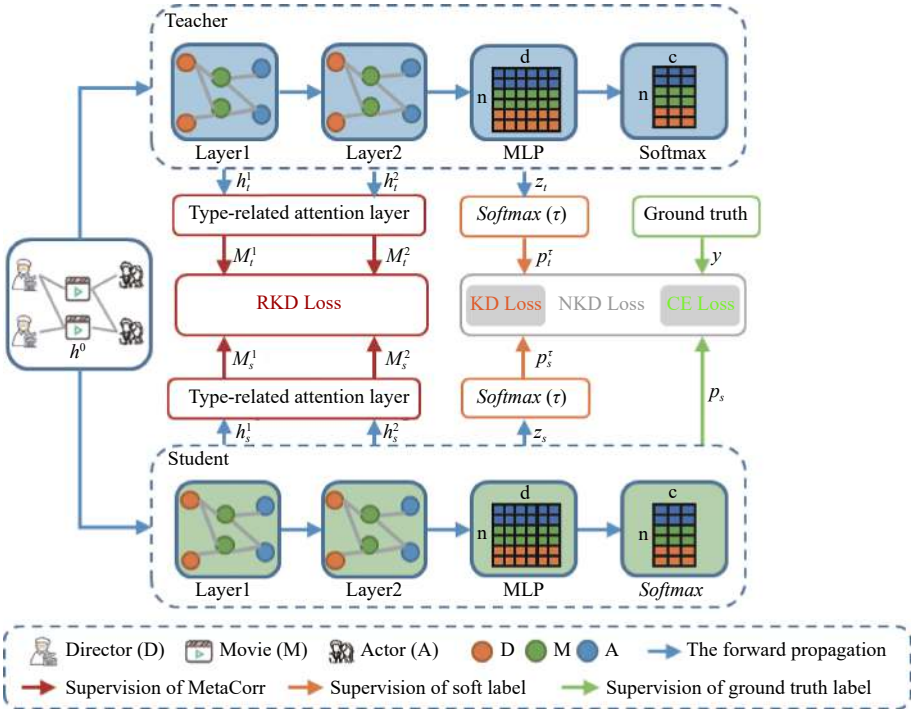


图 16 HIRE[114]的构造图知识蒸馏

综上, 基于构造图的图蒸馏方法在 GNN 上均取得了令人印象深刻的性能提升, 已成为当前 GNN 蒸馏学习新范式. 由于面向图神经网络的图蒸馏方法最近几年才被广大学者关注, 因此该领域还有很多问题需要探索和亟需解决, 具体可参见第 8 节展望部分.

5 基于图知识的模型自蒸馏方法

随着图知识蒸馏的发展, 还有一类模型自蒸馏方法被提出, 这极大地吸引了学者广泛的关注, 成为当前研究的热点之一. 因此, 本节将重点针对图神经网络模型中的自蒸馏方法进行汇总, 将基于图知识的模型自蒸馏方法依据蒸馏位置归类为输出层、中间层和构造图这 3 类知识, 具体细分见表 4. 同样地, 这里仅重点介绍各类图蒸馏算法最显著的知识蒸馏形式, 即若同时具有输出层、中间层知识的蒸馏方法默认划分为中间层知识蒸馏方法, 若同时具有这 3 类知识的蒸馏方法则默认划分为构造图知识蒸馏方法.

表 4 基于图知识的模型自蒸馏方法汇总

方法	蒸馏位置			距离度量	任务	应用
	输出层	中间层	构造图			
LinkDist[115]	√	—	—	MSE	模型压缩	节点分类
IGSD[116]	√	—	—	InfoCE	图级任务	图分类, 分子性质预测
GNN-SD[118]	—	√	—	KL, L2	缓解过平滑	节点、图分类
SDSS[119]	√	√	√	KL, MSE	半监督学习	多任务节点分类
SAIL[117]	√	—	—	KL	无监督学习	节点分类, 节点聚类, 链接预测

● 相同点: 针对基于图知识的模型自蒸馏方法中基于输出层、中间层和构造图这 3 类知识蒸馏形式, 每类图知识蒸馏算法的共性在于它们均是基于相同位置处知识的提取.

● 不同点: 针对每类图知识蒸馏算法, 它们的差异性体现在多个方面, 如具体方法实现、距离度量函数、下游任务及应用等方面上, 具体参见表 4. 如在图知识蒸馏中用 KL 散度的, 基于输出层有 SAIL[117], 基于中间层有 GNN-SD[118], 基于构造图有 SDSS[119]. 具体地, 在输出层知识进行蒸馏的方法中: LinkDist[115]采用 MSE 距离度量方式将图知识蒸馏应用在模型压缩的节点分类应用场景中, IGSD[116]利用 MSE 度量方式对学生模型压缩并用于节点分类任务上, SAIL[117]则使用 KL 将图知识蒸馏用于图无监督学习的常见下游节点分类、节点聚类和链接预测上; 基于中间层知识进行蒸馏的 GNN-SD[118]方法使用 KL 和 L2 距离度量方式用于下游图级任务的图分类和分子性质预测上; 基于构造图知识进行蒸馏的 SDSS[119]则利用 KL 和 MSE 结合的距离度量方式将图知识蒸馏用于下游半监督学习的多任务节点分类中.

5.1 基于输出层知识

基于输出层知识同样也是自蒸馏最基础且最常用的方法, 本质上就是提取预训练教师模型所蕴含的标签类别相关知识. 在图学习任务中, 尤其图的半监督甚至自监督学习中, 面临着标签数据难以获取这一大难题. 基于此, 研究人员开始尝试将模型自蒸馏方法应用在 GNN 上. 其中, 最具有代表性的工作就是 Zhang 等人[116]在 2020 年提出的一种基于自蒸馏的图级表示学习框架 IGSD, 该框架通过对图实例的增广视图进行实例判别来迭代地执行师生蒸馏. IGSD 的自蒸馏架构如图 17 所示.

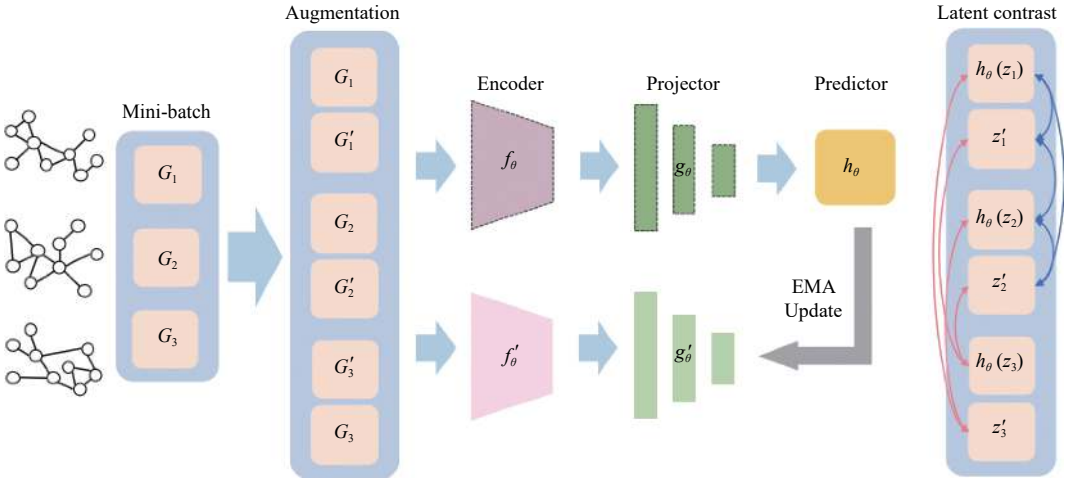


图 17 IGSD 的输出层知识蒸馏[116]

与传统的知识蒸馏不同, IGSD 用学生模型的指数移动平均值构建教师模型, 并提取自身的知识, 通过图增广视图迭代地执行教师与学生之间的蒸馏. 本质就是在对比学习的框架下, 结合自蒸馏技术使得教师网络和学生网络同时进行训练, 进而增强图的表示能力. 同时, 该工作还将 IGSD 扩展到半监督学习场景, 通过联合使用监督和自监督的对比损失来优化 GNN. 最后, IGSD 在半监督图分类和分子性质预测任务中超越了最先进的方法, 并在自监督图分类任务中实现了与最先进的方法相当的性能.

随后, Luo 等人[115]提出 LinkDist 自蒸馏方法, 旨在通过从图网络的边中蒸馏知识, 使得 MLP 在图分类任务上达到了甚至超过了 GCN 模型表现力. 此外, 作者还从随机点对中蒸馏“逆边”的知识, 进一步提升了模型的效果. IGSD 和 LinkDist 这两项工作, 很好地验证了输出层知识方法的优越性, 表明基于图知识模型自蒸馏的输出层知识蒸馏方式在 GNN 上的巨大潜力.

5.2 基于中间层知识

基于输出层知识虽然简单有效, 但仅依靠 GNN 模型输出类别标签这一类监督信息的话, 模型的表达能力有限. 于是, 研究者开始探索 GNN 中间图卷积层位置模型蕴含的丰富信息, 希望挖掘到更具节点特征表达能力的知

识, 代表性方法有 GNN-SD[118]和 SAIL[117].

如图 18 所示, 考虑到两阶段 T-S 蒸馏方法训练耗时且学生模型的性能受限于教师模型的选取, Chen 等人[118]提出 GNN-SD 方法来替代传统两阶段的 GNN 蒸馏方法, 将 GNN 中间层知识从浅层蒸馏到深层以缓解 GNN 面临的过平滑问题, 同时显著增强 GNN 的性能.

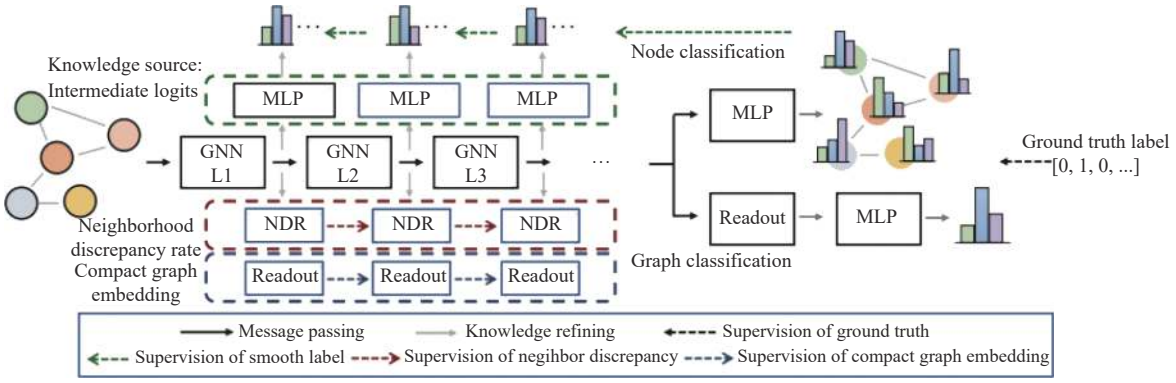


图 18 GNN-SD 的中间层知识蒸馏[118]

具体地, 该方法提出邻域差异率 (neighborhood discrepancy rate, NDR) 指标, 用于量化图嵌入浅层的非光滑性, 将此作为知识提炼到 GNN 的深层表示中. 通过这种方法, 模型通过自身的蒸馏策略保持了从初始嵌入图到最终嵌入输出的非光滑性. 同时, 在 NDR 的基础上, GNN-SD 还设计了自适应偏差保留 (adaptive discrepancy retaining, ADR) 正则化器, 以增强知识的可转移性, 使知识在 GNN 层之间保持较高的邻域偏差. 通过在多个流行的 GNN 模型上进行实验, GNN-SD 方法的有效性和泛化能力得到了验证, 确实可以有效地缓解 GNN 过平滑问题, 同时还能够大大降低两阶段知识蒸馏的训练成本.

另一个具有代表性的工作是, Yu 等人[117]在 2022 年提出的 GNN 上第 1 个通用的自监督蒸馏框架 SAIL, 其模型框架图见图 19.

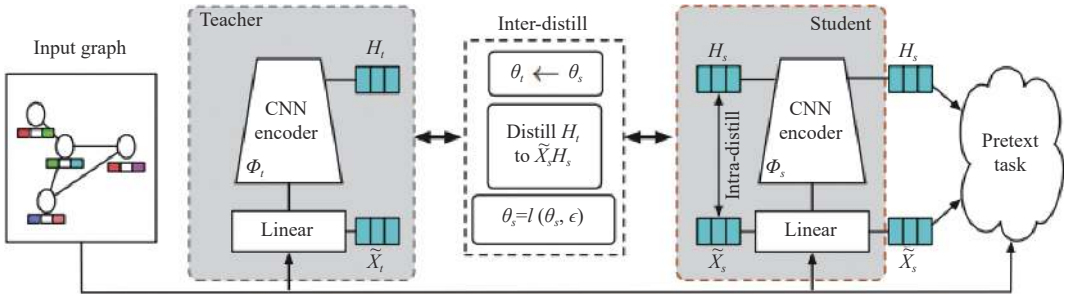


图 19 SAIL 的中间层知识蒸馏[117]

在这项工作中, 作者发现 GNN 的性能取决于节点特征的平滑性和图结构的局部性. 于是, SAIL 设计了两个自蒸馏正则化模块来平滑图拓扑和节点特征的近似度差异. 从图 19 可以得知, SAIL 框架主要包含两个互补的自蒸馏模块, 即 Intra-distill 和 Inter-distill 蒸馏模块, 通过 Intra-distill 和 Inter-distill 来迭代的利用 GNN 中间层平滑节点特征来修正 GNN 浅层表示. 为了评估学习到的节点表示能力, SAIL 进行了丰富的实验, 包括节点分类、节点聚类和链接预测任务. 实验结果表明, SAIL 有助于学习具有强竞争力的浅层 GNN, 优于当前有监督或无监督方式训练得到的 GNN.

GNN-SD 和 SAIL 这两项工作的初步研究成果, 揭示了一种很有前途的实现 GNN 自蒸馏的方法.

5.3 基于构造图知识

除了输出层和中间层知识蒸馏形式外, 研究人员也关注到了基于构造图的知识蒸馏形式. 相比于前面两种知

识, 构造图知识可以提取到模型中节点特征之间的关联信息, 可以充分挖掘并利用模型知识, 为基于图知识模型自蒸馏的研究打开了新的视角. 图 20 展示了在 GNN 上利用基于构造图知识进行自蒸馏的训练过程.

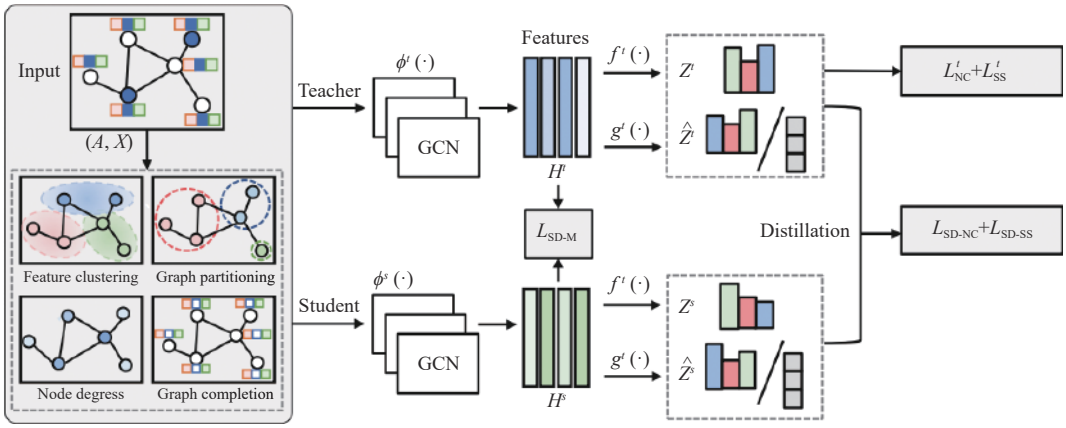


图 20 SDSS 的构造图知识蒸馏[119]

Ren 等人[119]认为图结构和标签之间的不匹配影响模型的性能, 提出了一个多任务的自蒸馏框架 SDSS, 将自监督学习和自蒸馏注入图卷积网络中, 通过挖掘图和标签中的信息, 解决结构端和标签端不匹配问题. 具体地, SDSS 方法利用 4 个基于前置任务的自监督学习来提取不同层次的图相似度信息, 以促进图卷积网络的局部特征聚合. 实验结果表明, 该方法在几种经典的图卷积模型上都取得了令人印象深刻的性能提升. SDSS 方法的成功, 表明自监督和自蒸馏在 GNN 框架中得到了很好的结合, 这为之后的研究人员提供了新的思路和方向.

6 实验分析

本节对面向深度神经网络的图知识蒸馏方法、面向图神经网络的图知识蒸馏方法和基于图知识的模型自蒸馏方法分别展开实验进行对比分析. 首先, 本节介绍常用的数据集; 然后对实验设置进行介绍; 最后对图知识蒸馏方法的实验结果进行分析.

6.1 数据集

为了对比面向深度神经网络的图知识蒸馏方法前后的实验效果, 本文选用深度神经网络中常用的 2 个数据集, 包括 CIFAR-10[120]和 CIFAR-100[120]. 数据集具体信息如表 5 所示.

表 5 面向深度神经网络的图知识蒸馏领域常用数据集

数据集	图像数	训练集	测试集	类别数
CIFAR-10	60000	50000	10000	10
CIFAR-100	60000	50000	10000	100

为了对比面向图神经网络的图知识蒸馏方法前后的实验效果, 本文选用图神经网络中常用的 7 个数据集, 包括 Cora、CiteSeer、PubMed、Amazon-Photo (A-P)、Amazon-Computers (A-C)、Coauthor-Physics (Physics)、Coauthor-CS (CS). 数据集具体信息如表 6 所示.

CIFAR-10: 是一个小型彩色图像数据集, 共有 60000 张彩色图像, 一共分为 10 个类, 每类 6000 张图. 其中, 这里面有 50000 张图像用于训练, 另外 10000 图像用于测试.

CIFAR-100: 由 100 个类的 60000 个彩色图像组成, 每个类包含 600 个图像. 其中, 每类各有 500 个训练图像和 100 个测试图像.

Cora: 是一个由机器学习论文构成的一个基准引文数据集[121]. 节点表示论文, 边表示引用关系. 每个节点都有一个 1 433 维的特征, 类标签表示每篇论文所属研究领域, 任务是根据引文网络将论文分类为不同的领域.

CiteSeer: 是另一个常用的基准引文数据集[121]. 其中, 每个节点代表一篇论文, 每条边代表两篇论文之间的引用关系, 节点特征维度是 3 703 维, 类标签有 6 种, 任务是预测某出版物的所属类别.

PubMed: 也是一个引文网络[121], 包含 19 717 个节点, 44 324 条边. 其中节点表示与糖尿病相关的文章, 边表示引用文章间的关系. 节点特征为 TF/IDF 加权词频, 有 500 维. 类别标签有 3 类, 任务是预测论文的糖尿病类型.

A-P 和 A-C: 是亚马逊的一个商品购买网络[122]. 节点表示商品, 连边表示两种经常被一起购买. 节点特征用商品评论的词袋表示, 任务是预测商品的类别.

Physics 和 CS: 是常用的引用网络, 是从 2016 年 KDD 杯挑战赛的微软学术合著图提取出来的[122]. 节点表示作者, 边表示作者之间是否是合作关系. 节点特征用每个作者发表论文的关键词表示, 类别标签表示每个作者的研究领域. 给定每个作者论文的关键词, 任务是将作者划分到他们各自的研究领域.

表 6　面向图神经网络的图知识蒸馏领域常用数据集

数据集	节点数量	边数量	特征维度	类别数
Cora	2 708	5 278	1 443	7
CiteSeer	3 327	4 552	3 703	6
PubMed	19 717	44 324	500	3
A-P	7 650	119 043	745	8
A-C	13 752	245 778	767	10
Physics	34 493	247 962	8 415	5
CS	18 333	81 894	6 805	15

6.2　实验设置

面向深度神经网络的图知识蒸馏: 为了实验的简洁性, 本文选取具有代表性的 ResNet-20[123]模型作为深度神经网络模型框架, 测试图知识蒸馏方法在上述 2 个数据集图像分类任务上的表现. 其中, 在分类指标上, 本文选取的是 Accuracy 指标. 在知识蒸馏方法的选取上, 本文使用的是经典的 KD 和 IRG、RKD 以及 CC 这 3 种常用的图知识蒸馏方法. 具体分类效果可见表 7.

表 7　面向深度神经网络的图知识蒸馏方法 Accuracy 效果对比

模型	CIFAR-10	CIFAR-100
Teacher	0.923 7	0.689 2
+KD	**0.933 0**	<u>0.703 6</u>
+IRG	0.927 7	**0.703 7**
+RKD	0.927 2	0.694 8
+CC	<u>0.930 1</u>	0.692 7

注: 加粗部分表示最佳性能, 下划线部分表示次优性能

面向图神经网络的图知识蒸馏: 在节点分类任务的对比实验中, 本文选用最具代表性的图神经网络模型 (即 GCN[17]、GAT[32]、和 SAGE[29]), 在上述 7 个数据集上进行节点分类对比实验. 其中, 在知识蒸馏方法的选取上, 本文使用的是经典的 KD 和 CPF 图蒸馏方法. 在分类指标上, 本文选取的是 F1-Micro 和 F1-Macro 指标. 具体分类效果可见表 8–表 10. 同样地, 在聚类任务上, 本文采用 NMI 和 ARI 聚类指标, 使用 GCN、GAT 和 SAGE 这 3 个图神经网络模型在 Cora 等 7 个数据集上应用 KD 和 CPF 知识蒸馏方法, 具体实验结果见第 6.3 节. 此外, 为了进一步定量分析知识蒸馏效果, 本文还给出了节点可视化实验结果. 具体地, 就是对 GCN、GAT 和 SAGE 模型知识蒸馏前后的节点表示进行 t-SNE 降维, 可视化结果见图 21–图 23.

表 8 GCN 教师模型及其学生变体模型节点分类效果

模型	指标	Cora	CiteSeer	Pubmed	A-P	A-C	Physics	CS
GCN	*F*1-Micro	0.8096	0.7086	0.7912	0.6441	0.4727	0.9226	0.8918
	*F*1-Macro	0.7985	0.6789	0.7862	0.6374	0.4347	0.9002	0.8698
+KD	*F*1-Micro	0.8276	0.7364	**0.8032**	0.7982	0.6759	0.9322	0.9105
	*F*1-Macro	0.8163	**0.6916**	**0.7964**	0.7859	0.6672	0.9107	**0.8862**
+CPF	*F*1-Micro	**0.8562**	**0.7530**	0.6811	**0.9313**	**0.8459**	**0.9476**	**0.9126**
	*F*1-Macro	**0.8261**	0.6623	0.6716	**0.9162**	**0.8456**	**0.9304**	0.8860

注: 加粗部分表示最佳性能, 下划线部分表示次优性能

表 9 GAT 教师模型及其学生变体模型节点分类效果

模型	指标	Cora	CiteSeer	PubMed	A-P	A-C	Physics	CS
GAT	*F*1-Micro	0.8208	0.7032	0.7722	0.8054	0.6559	0.9252	0.9066
	*F*1-Macro	0.8116	0.6732	0.7684	0.7972	0.6244	0.9023	0.8824
+KD	*F*1-Micro	0.8410	0.7264	0.7848	0.8421	0.6688	0.9350	**0.9090**
	*F*1-Macro	**0.8331**	**0.6818**	0.7780	0.8273	0.6535	0.9136	**0.8859**
+CPF	*F*1-Micro	**0.8576**	**0.7541**	**0.7949**	**0.9158**	**0.8456**	**0.9407**	0.9031
	*F*1-Macro	0.8295	0.6692	**0.7938**	**0.8981**	**0.8516**	**0.9219**	0.8743

注: 加粗部分表示最佳性能, 下划线部分表示次优性能

表 10 SAGE 教师模型及其学生变体模型节点分类效果

模型	指标	Cora	CiteSeer	PubMed	A-P	A-C	Physics	CS
SAGE	*F*1-Micro	0.7980	0.7052	0.7844	0.8754	0.7660	0.9239	0.9203
	*F*1-Macro	0.7862	0.6769	0.7824	0.8663	0.7599	0.9021	0.8999
+KD	*F*1-Micro	**0.8198**	0.7202	0.7962	0.8853	**0.7838**	**0.9379**	**0.9262**
	*F*1-Macro	**0.8100**	**0.6877**	0.7935	0.8763	0.7821	**0.9163**	**0.9062**
+CPF	*F*1-Micro	0.8183	**0.7365**	**0.8053**	**0.9249**	0.7824	0.9363	0.9016
	*F*1-Macro	0.7861	0.6148	**0.8051**	**0.9067**	**0.8078**	0.9135	0.8682

注: 加粗部分表示最佳性能, 下划线部分表示次优性能

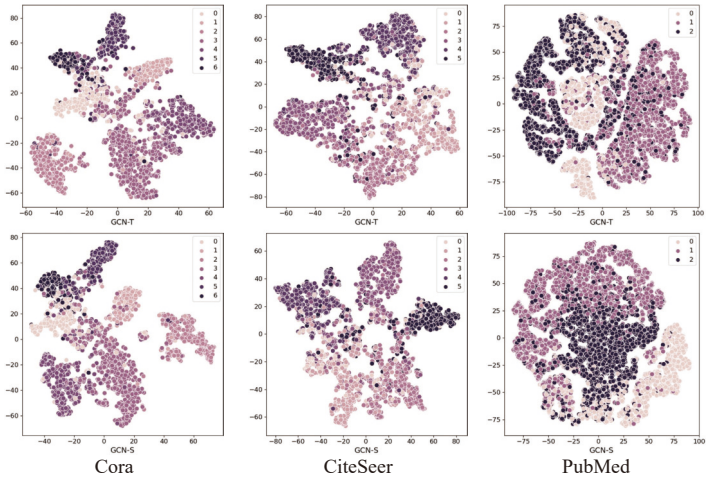


图 21 GCN 教师模型及其学生变体节点可视化效果

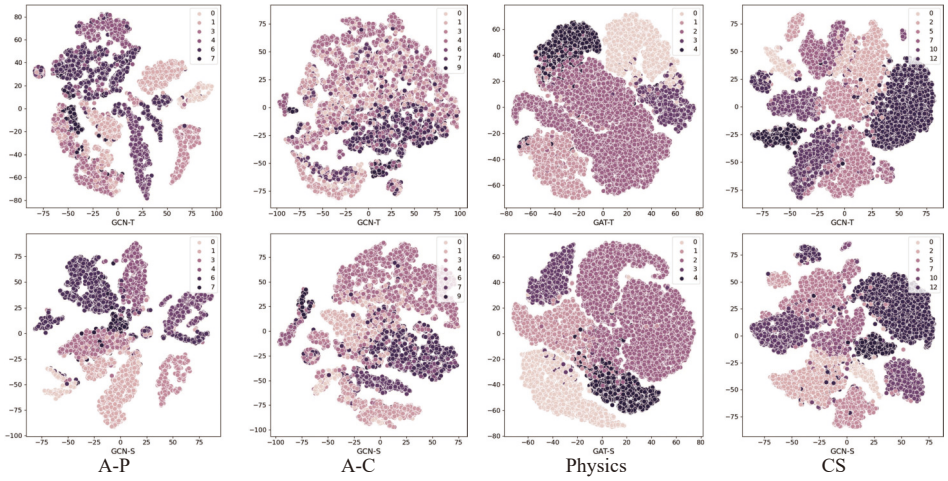


图 21 GCN 教师模型及其学生变体节点可视化效果 (续)

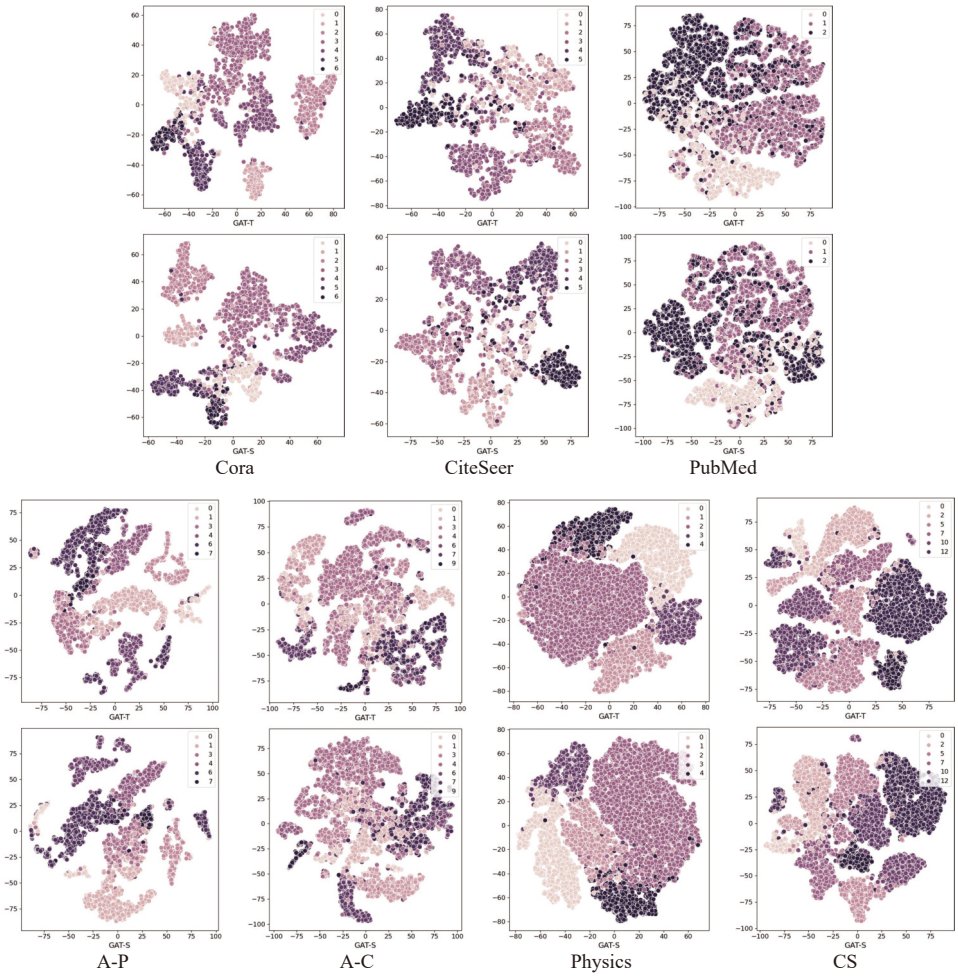


图 22 GAT 教师模型及其学生变体节点可视化效果

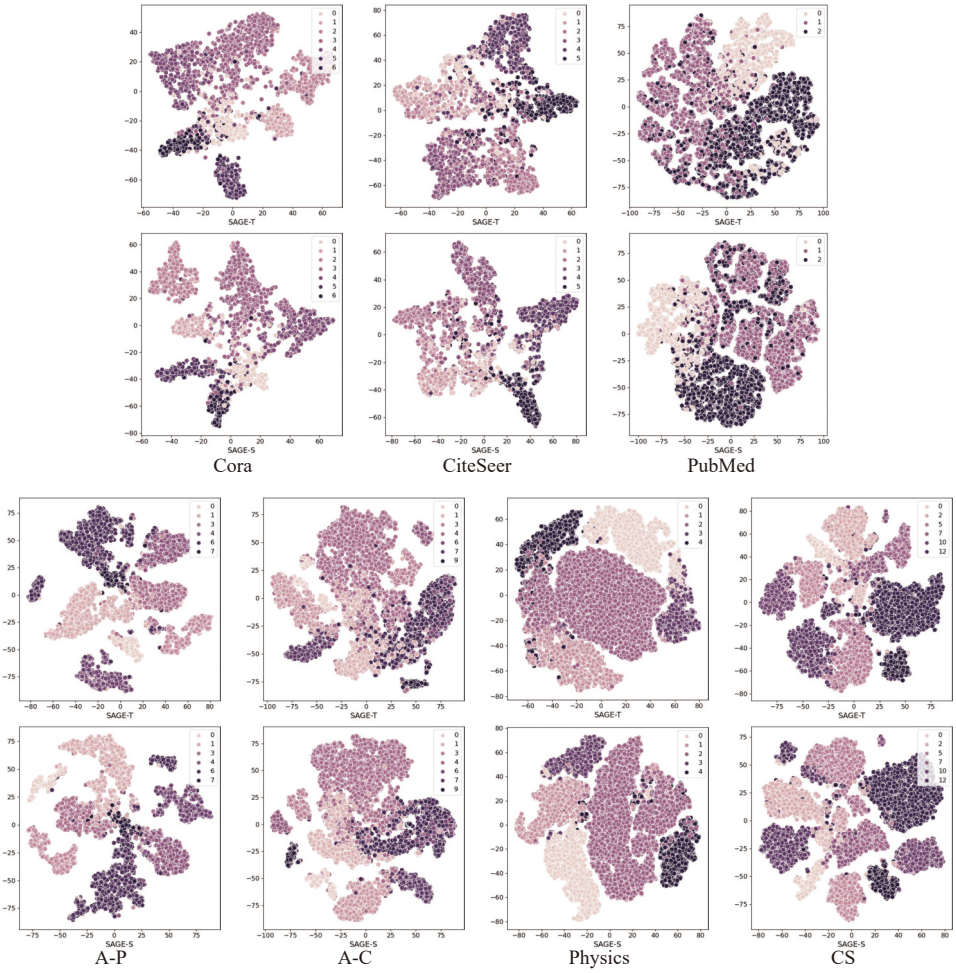


图 23 SAGE 教师模型及其学生变体节点可视化效果

基于图知识的模型自蒸馏方法实验: 为了实验的简洁性, 本文选取经典的 GCN 模型作为图神经网络模型框架, 以节点分类作为任务, 在具有代表性的 Cora、CiteSeer 和 PubMed 这 3 个数据集上测试图知识蒸馏效果. 其中, 在分类指标上, 本文选取的是 Accuracy 指标. 在知识蒸馏方法的选取上, 本文使用的是经典的 KD 和 LinkDist、SAIL 以及 SDSS 这 3 种模型自蒸馏方法. 具体分类效果可见表 11.

表 11 基于图知识的模型自蒸馏方法 Accuracy 效果对比

模型	Cora	CiteSeer	PubMed
Teacher	0.8183	0.6762	0.7859
+KD	*0.8005*	0.6821	*0.7571*
+LinkDist	*0.7572*	0.7119	*0.7484*
+SAIL	0.8463	0.7424	**0.8381**
+SDSS	**0.8600**	**0.7613**	0.8221

注: 加粗部分表示最佳性能, 下划线部分表示次优性能, 斜体部分表示性能下降

本文所做的全部实验都运行在 NVIDIA Tesla 的 V100 GPU 上, 并基于 PyTorch-1.6.0 版本为 0.6 的 DGL 图学习库实现.

6.3 实验结果与分析

● 面向深度神经网络的图知识蒸馏实验结果分析. 从前文表 7 可以得知, KD 和 IRG、RKD 以及 CC 这 3 种图知识蒸馏方法均一致显著提升了 ResNet-20 教师模型的图像分类效果. 其中, KD 在 CIFAR-10 的蒸馏表现最佳, CC 表现次之. 在 CIFAR-100 数据集上: IRG 的蒸馏效果最好, 将教师模型的图像分类性能从 0.698 2 提升到了 0.703 7; KD 的表现次之, 将教师模型的图像分类性能从 0.698 2 提升到了 0.703 6. 尽管 KD 和 IRG、RKD 以及 CC 这 3 种图知识蒸馏方法对 ResNet-20 模型的增益效果不尽相同, 但是他们的性能相差不大, 表现相当. 这一定程度上反映了深度神经网络和知识蒸馏算法的结合工作对 ResNet-20 模型的提升到达了一定的瓶颈. 可以探索新的图蒸馏学习范式进一步提升面向深度神经网络的图知识蒸馏方法的蒸馏效果, 如与对抗学习、神经架构搜索、图神经网络等新技术结合, 这部分的讨论具体可见第 8 节 (5) 图蒸馏学习新范式.

● 面向图神经网络的图知识蒸馏节点分类实验结果分析. 由前文表 8–表 10 节点分类结果可以看出, 无论在经典 KD 还是 CPF 图蒸馏的指导下, GCN、GAT 和 SAGE 学生变体在 Cora 等 7 个数据集上的分类性能均获得了一致显著的提升. 尤其是通过表 8 和表 9, 可以明显发现在 GCN 和 GAT 模型框架下, CPF 的蒸馏效果远好于经典 KD 蒸馏效果. 但是 SAGE 的学生变体模型表现相反, 尽管在 SAGE 模型框架下, 大多数情况下 KD 蒸馏效果好于 CPF, 但它们相差不多 (+KD≈+CPF), 表现相当. 此外, 尽管 CPF 是多个位置知识蒸馏的集成方法, 但是该图蒸馏方法在一些数据集上的表现不如 KD, 如 GCN 模型在 PubMed 数据集上的蒸馏表现, GAT 模型在 CS 数据集上的蒸馏表现以及 SAGE 模型在 Cora、Physics 和 CS 数据集上的蒸馏表现均低于 KD 蒸馏效果. 综上, 虽然 KD 和 CPF 图知识蒸馏方法均一致显著提升了 GNN 模型的节点分类性能, 但它们无法在 Cora、CiteSeer、PubMed、A-P、A-C、Physics 和 CS 这 7 个数据集上的性能同时保持最佳. 同样地, 同一种图知识蒸馏算法在不同的 GNN 模型上的蒸馏效果不尽相同. 例如 CPF 图蒸馏算法可以将 GCN 性能从 80.6% 提升到 85.62%, 将 GAT 模型从 82.08% 提升到 85.76%, 将 SAGE 模型从 79.80% 提升到 81.83% (这里以节点分类 $F1$-Micro 指标为例). 这反映了, 在设计图知识蒸馏算法时, 不仅要考虑不同蒸馏位置知识的组合性, 而且要考虑 GNN 模型的适用性, 还需要考虑数据集的普适性. 也就是说, 需要设计一种可以适用于任意图神经网络模型的强大图知识蒸馏算法, 同时可以在任意的数据集上获得 SOTA (state-of-the-art) 蒸馏效果.

● 面向图神经网络的图知识蒸馏节点聚类实验结果分析. 除了执行节点分类任务外, 本文还针对 GCN 等 3 个模型在 Cora 等 7 个数据集上进行了节点聚类实验. 具体的聚类效果见表 12–表 14.

表 12 GCN 教师模型及其学生变体模型节点聚类效果

模型	指标	Cora	CiteSeer	PubMed	A-P	A-C	Physics	CS
GCN	NMI	0.556 8	0.429 1	0.371 1	0.423 5	0.339 9	0.702 9	0.673 6
	ARI	0.512 0	0.424 1	0.409 4	0.261 9	0.208 3	0.680 6	0.533 6
+KD	NMI	**0.601 1**	0.465 5	**0.387 4**	**0.593 2**	0.457 8	0.711 1	**0.714 5**
	ARI	**0.593 3**	0.462 0	**0.440 1**	**0.465 5**	0.288 3	0.689 0	**0.602 5**
+CPF	NMI	0.598 8	**0.471 4**	*0.193 5*	0.588 8	**0.529 9**	**0.751 9**	*0.529 9*
	ARI	0.580 1	**0.472 1**	*0.121 4*	0.387 2	**0.298 5**	**0.822 8**	*0.298 5*

注: 加粗部分表示最佳性能, 下划线部分表示次优性能, 斜体部分表示性能下降

表 13 GAT 教师模型及其学生变体模型节点聚类效果

模型	指标	Cora	CiteSeer	Pubmed	A-P	A-C	Physics	CS
GAT	NMI	0.605 6	0.429 7	0.362 6	0.654 5	0.497 5	0.766 9	0.753 1
	ARI	0.563 4	0.425 7	0.391 0	0.531 1	0.401 8	0.839 1	0.688 9
+KD	NMI	**0.614 5**	0.455 0	0.375 4	**0.681 4**	**0.556 7**	**0.771 1**	**0.771 9**
	ARI	**0.579 9**	**0.444 9**	0.416 9	**0.597 5**	**0.476 7**	**0.850 6**	**0.793 0**
+CPF	NMI	0.606 6	**0.455 1**	**0.402 1**	*0.511 3*	0.498 1	*0.614 7*	*0.585 0*
	ARI	*0.510 9*	*0.417 7*	**0.426 6**	*0.288 4*	*0.299 4*	*0.565 4*	*0.437 1*

注: 加粗部分表示最佳性能, 下划线部分表示次优性能, 斜体部分表示性能下降

表 14 SAGE 教师模型及其学生变体模型节点聚类效果

模型	指标	Cora	CiteSeer	PubMed	A-P	A-C	Physics	CS
SAGE	NMI	0.5707	0.4374	0.4083	0.6870	0.5380	0.7641	0.7988
	ARI	0.5433	0.4457	0.4564	0.5813	0.3686	0.8238	0.7509
+KD	NMI	**0.5921**	0.4618	**0.4177**	**0.7010**	**0.5775**	**0.7854**	**0.8149**
	ARI	**0.5825**	0.4597	**0.4632**	**0.6175**	**0.4499**	**0.8643**	**0.8397**
+CPF	NMI	*0.4892*	**0.4737**	*0.3598*	*0.4666*	*0.4808*	*0.6323*	*0.5779*
	ARI	*0.2965*	**0.4805**	*0.3724*	*0.2803*	*0.3104*	*0.5655*	*0.3894*

注: 加粗部分表示最佳性能, 下划线部分表示次优性能, 斜体部分表示性能下降

从表 12–表 14 可以得知, KD 和 CPF 带给 GNN 教师模型的增益效果相差较大. 整体上, 经过 KD 知识蒸馏, GCN、GAT 和 SAGE 的学生模型聚类性能均得到了一定的性能提升, 但是它们的蒸馏效果表现不尽相同. 其中, 从表 12 得知: 在 Cora、PubMed、A-P 和 CS 数据集上 KD 效果远大于 CPF, 然而在 CiteSeer、A-C 和 Physics 上 CPF 效果好于 KD. 从表 13 和表 14 可以观察出: 在 GAT 和 SAGE 模型骨架下, KD 效果整体好于 CPF 甚至于可以使得对应学生模型在各个数据集下的性能保持最佳.

另外, 本文发现图知识蒸馏实验中的一个有趣的实验现象, CPF 图知识蒸馏算法会损害 GNN 模型性能. 特别地, 在 GAT 和 SAGE 模型框架下, CPF 在 Cora 等 7 个数据集上反而会降低其对应教师模型的节点聚类性能. 例如, 经过 CPF 蒸馏, 基于 SAGE 模型框架下的 CS 数据集的教师性能从 79.88% 降低到 57.79% (这里以节点聚类 NMI 指标为例), 这和 CPF 图知识蒸馏算法在节点分类上表现大相径庭. 这一定程度上说明 GNN 和图蒸馏算法的结合工作仍有很大挑战, 如何更好地将图知识蒸馏应用在 GNN 上仍旧需要进一步的探索. 关于这部分, 本文在第 8 节进行了深度讨论和展望.

● 面向图神经网络的图知识蒸馏节点可视化实验结果分析. 除了进行节点分类和聚类定量分析图知识蒸馏效果外, 本文还进行了节点可视化定性分析. GCN、GAT 和 SAGE 变体模型的节点表征经过 t-sne 算法降维后的可视化结果如图见图 21–图 23. 其中, 第 1 行是对应教师模型节点可视化效果, 第 2 行是对应学生模型在图蒸馏下的可视化结果. 通过这 3 个图, 可以清晰发现经过知识蒸馏, GCN、GAT 和 SAGE 的学生聚类效果均得到了改善, 不同种类节点间的边界间隔变大, 相同种类的节点聚拢更加紧密. 尽管图知识蒸馏算法可以提升 GNN 模型的节点表示能力, 使得不同类别标签的分类界面变得更加清晰, 但是它们在不同的数据集上的效果不尽相同. 这说明当前图知识蒸馏在 GNN 上的研究仍有巨大的潜力, 还有很多问题值得进一步研究和探索, 关于这部分我们在后面第 8 节部分进行了深度的讨论和展望.

● 基于图知识的模型自蒸馏实验结果分析. 由前文表 11 节点分类结果可以看出, 无论在经典 KD 还是 LinkDist、SAIL 和 SDSS 这 3 种模型自蒸馏方法图蒸馏的指导下, GCN 学生变体在 Cora 等 3 个数据集上的分类性能或多或少得到了一定的性能提升. 特别地, SAIL 和 SDSS 的蒸馏效果远好于经典 KD 蒸馏效果, 且大幅度提升了 GCN 教师模型的性能. 但是 LinkDist 的表现不尽相同, 其蒸馏表现在 Cora 和 PubMed 数据集上低于 KD 的蒸馏效果. 同时, 本文还发现 KD 和 LinkDist 在 Cora 和 PubMed 数据集上反而会降低其对应教师模型的节点分类性能. 综上, 虽然 KD 和模型自蒸馏方法可以提升图神经网络模型的节点分类性能, 但它们无法在 Cora、CiteSeer 和 PubMed 这 3 个数据集上的性能同时保持最佳. 这同样反映了, 在设计图知识蒸馏算法时, 还需要考虑蒸馏方式选择的适当与否, 同时蒸馏位置和距离度量的函数选择也影响着蒸馏效果, 有关这部分的讨论具体可见第 8 节. 此外, 基于图知识的模型自蒸馏方法目前处于初步探索阶段, 研究方法鲜少且缺乏理论支撑, 需要充分探索图知识蒸馏方法背后的蒸馏数学原理机制, 从而设计出高效图知识蒸馏方法, 有关这部分的讨论具体可见第 8 节 (4) 可解释性理论分析.

综上, 这 3 类方法图知识蒸馏方法凭借其模型压缩、模型增强、简单高效等优势可以提升 CNN/GNN 模型性能, 成功应用在推荐系统等实际应用场景中. 尽管这些方法取得了不错的成效, 但它们仍然存在一定的不足. 本文接下来在第 8 节对图知识蒸馏的可改进方向进行了展望: (1) 蒸馏位置的确定; (2) 蒸馏方式的选择; (3) 距离度量的函数选取; (4) 可解释性理论分析; (5) 图蒸馏学习新范式.

7 应 用

知识蒸馏自提出以来, 受到了学术界和工业界的大量关注. 随着知识蒸馏技术的发展, 图知识蒸馏在模型压缩和模型增强等方面均取得了优异的表现, 而且在计算机视觉、自然语言处理、推荐系统等领域有着非常重要的应用和广阔的前景. 在本节中, 总结了图知识蒸馏常见的几个应用场景, 这对更好地理解和使用图知识蒸馏技术至关重要 (比如大多采用 T-S 蒸馏方式; KD 和 GNN 等技术的结合), 也是未来值得关注的研究工作.

7.1 计算机视觉

图知识蒸馏作为一种有效的模型压缩/模型增强技术, 广泛应用在人工智能的不同领域, 尤其是在计算机视觉 (computer vision) 领域中. 近年来, 各种各样的图知识蒸馏算法被提出应用于不同的视觉任务中. 其中, 图知识蒸馏主要应用在图像分类[58,68,71−73,76]下游任务中, 实现模型增强、可解释性、模型压缩等目标. 图知识蒸馏在图像识别[57,63,69]上, 也有着非常重要的应用, 通过构造辅助图作为知识的载体, 挖掘教师模型中样本间的关系知识传递到学生模型中, 进一步提升学生模型的性能. 此外, 在无监督学习场景下, 图知识蒸馏方法也被用于计算机视觉领域中的行人重识别[59,81]建模问题中. 另外, 如表 15 所示, 图知识蒸馏还可以用于目标检测[70,80]、机器人定位[66]、视频分类[78]、事件预测[98]和道路标记[83]等视觉任务中.

表 15 图知识蒸馏应用领域总结表

应用领域	应用问题	蒸馏模型	相关论文
计算机视觉	图像分类	CNN	HKD[68], GKD[76], SPKD[58], HKDIFM[71], KDExplainer[72], TDD[73]
	图像识别	CNN	KTG[63], MHGD[69], IRG[57]
	机器人定位	GNN	GCLN[66]
	目标检测	CNN	DOD[70], GD[80]
	视频分类	CNN	BAF[78]
	事件预测	GNN	EGAD[98]
	行人重识别	CNN	GCMT[81], CC[59]
	道路标记	CNN	IntRA-KD[83]
自然语言处理	视觉对话	CNN	CAG[75]
	关系抽取	CNN	DKWISL[62]
	视频字幕	CNN	SPG[65]
	机器翻译	CNN	LAD[79]
	度量学习	CNN	RKD[56]
推荐系统	增量学习	GNN	LWC-KD[96]
	协同过滤	CNN	DGCN[63]
	冷启动	GNN	PGD[101]
	尾部泛化	GNN	Cold Brew[100]
多任务学习	迁移学习	CNN	IEP[67]
	图表示增强	GNN	GRL[107]
	图像识别	CNN	MHGD[69]
	自蒸馏	GNN	SDSS[119]
零样本学习	无数据蒸馏	GNN	GFKD[85]
	模型增强	GNN	HGKT[109]

7.2 自然语言处理

自然语言处理 (natural language processing, NLP) 是计算机科学领域和人工智能领域中的一个重要分支, 是当前热点研究领域之一. NLP 模型发展日新月异, 从 RNN、Transformer、ELMo、GPT、BERT 再到如今的 GPT-3, 其模型结构和参数量变得越来越复杂且庞大, 这严重阻碍了语言模型的部署和训练. 知识蒸馏的出现, 提供了一种

有效的轻量化深度语言模型知识迁移方法, 可以简单高效地解决语言模型部署问题, 成为 NLP 领域的研究热点. 如今, 越来越多的图知识蒸馏工作被提出来处理 NLP 问题, 包括视觉对话[75]、机器翻译[79]、关系抽取[62]等任务, 具体见表 15.

7.3 推荐系统

推荐系统 (recommended system), 顾名思义, 就是根据用户的属性、历史行为等信息来建模用户的偏好, 进而产生用户喜欢的推荐. 随着深度学习的快速发展, 推荐系统的模型结构变得越来越复杂, 网络深度越来越深, 模型参数也变得越来越多. 同样地, 在推荐系统领域也面临着模型计算昂贵, 无法在移动端或嵌入式设备上运行的难题. 为了解决模型效果和响应速度之间的矛盾, 图知识蒸馏应运而生. 利用图知识蒸馏技术, 预训练的强大教师模型中的丰富知识可以被蒸馏到在线推荐的轻量化学生模型中, 实现增强推荐系统学生模型的泛化能力, 从而达到推荐模型轻松部署上线的目标. 此外, 图知识蒸馏和推荐系统的结合还可以用来解决冷启动[101]、尾部泛化[100]、增量学习[96]等问题. 前文表 15 列举了图知识蒸馏和推荐系统相结合的代表性工作.

7.4 多任务学习

图知识蒸馏技术除了在上述计算机视觉、自然语言处理和推荐系统上具有广泛应用外, 还和其他新兴技术如图神经网络、迁移学习结合使用, 用于下游多任务学习 (multi-task learning) 上. 具体地, Lee 等人[67]提出一种基于主成分分析的可解释嵌入过程 (IEP) 知识蒸馏方法, 以解释和理解深度神经网络模型嵌入表示的过程. Ma 等人[107]利用构造图来控制教师的知识转移, 通过多任务学习将基于网络理论的图度量作为辅助任务来学习更好的图表示. Lee 等人[69]使用多头注意将教师嵌入过程中的知识提取出来, 并通过多任务学习使学生模型具有关系归纳偏置能力. Ren 等人[119]则是将自蒸馏和多任务学习相结合, 提出了一个两阶段训练的多任务自蒸馏框架表 15 整理了目前图知识蒸馏在该任务上的部分工作, 可供研究人员参考.

7.5 零样本学习

同样地, 在零样本学习 (zero-shot learning) 领域, 图知识蒸馏也表现优异. 例如, Deng 等人[85]首次提出针对 GNN 量身定制的无数据蒸馏方法, 该方法通过使用多元伯努利分布从预训练的 GNN 建模图结构来进行知识迁移, 并引入梯度估计器来优化这个框架. Wang 等人[109]提出了一种基于异构图的知识转移方法 HGKT, 借助于构造的结构化异构图来表示数据之间的关系, 将知识从可见类转移到新的不可见类, 解决了可见类和不可见类实例分类的问题. 前文表 15 给出了这部分工作所使用的主流方法.

8 展 望

作为一种知识迁移技术, 图知识蒸馏凭借其模型压缩、模型增强、简单高效等优势提升深度神经网络和新兴的图神经网络的模型性能, 成功应用在推荐系统等实际业务场景中. 虽然图知识蒸馏取得了令人满意的性能表现, 成为当前热门的研究邻域, 但它仍然有很多需要注意的问题和值得进一步探索的方向. 针对目前图知识蒸馏方法的不足, 本节提出图知识蒸馏研究的几个潜在研究方向.

(1) 蒸馏位置的确定. 通过对图蒸馏工作的归纳分析, 现有大多数图蒸馏方法都是利用不同类型的知识源组合, 包括输出层、中间层、和构造图知识. 然而, 目前还不清楚哪个位置的知识起着重要的影响, 尤其对于中间层和构造图知识, 有的选择中间某一卷积层, 有的选择所有卷积层, 但是具体选择哪一层进行蒸馏, 目前鲜有研究. 如何设计出一种速度更快、更加通用、精度有保证且可以同时建模各种类型知识的图蒸馏模型仍存在挑战, 特别是分析出输出层知识、中间层知识和构造图知识这三者之间的关系如何, 它们是如何相互作用和影响的, 这对于图结构数据信息的合理利用和对知识的充分挖掘至关重要, 这是图知识蒸馏邻域未来研究的重点.

(2) 蒸馏方式的选择. 当前流行的两大图蒸馏方式有 T-S 蒸馏模式和自蒸馏模式. T-S 蒸馏方式因其灵活可控和易于操作, 适用于大规模复杂教师模型的模型压缩任务上. 模型自蒸馏方式因其结构简单和训练高效, 广泛应用在开销较大的下游实际业务场景中. 但是这两种蒸馏方式仍存在不足: T-S 操作复杂、训练耗时, 自蒸馏缺少理论支撑且局限于教师和学生模型性能相当的问题场景中. 然而, 目前缺乏对两者蒸馏方式的对比研究. 为此, 研究蒸

馏方式的选择如何影响 KD 的有效性, 以及如何设计出一个高效的蒸馏框架, 是十分必要的.

(3) 距离度量的函数选取. 图蒸馏的表现与训练损失中距离度量函数的选取密不可分. 因为知识蒸馏是从教师模型中提取知识蒸馏到学生模型中, 该知识迁移的效果好坏体现在模型训练中损失函数的设计上, 即只能通过评估学生模型与教师模型中节点/节点间特征的接近程度来展现. 于是, 设计出一个好的图蒸馏损失函数至关重要. 然而, 损失函数选取方法多样, 有 KL、MSE、InfoCE 等, 对于在图蒸馏过程中具体选择哪一种损失函数以更好指导学生模型训练过程, 尚无定论. 因此, 如何根据具体场景和问题选取合适的距离度量函数成为图蒸馏技术中亟待解决的问题.

(4) 可解释性理论分析. 尽管已有大量的图知识蒸馏工作被成功应用在各种实际业务场景中, 但对知识蒸馏的可解释性理论分析仍然较少. 最近, 对知识蒸馏的可解释性已有一些初步的尝试, 如 Yuan 等人[48]从标签平滑角度解释了 KD 的原理, 认为 KD 的成功并不完全是因为教师类别之间的相似性信息, 而是由于软目标的正则化. 然而, 该发现只适用于分类任务, 并不适用于没有标签的任务[124]. Cheng 等人[125]从量化知识的角度来解释知识蒸馏, 即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. Mobahi 等人[126]通过在希尔伯特空间对训练数据的拟合, 首次证明了自蒸馏起着 L2 正则化器作用, 从而为自蒸馏方法提供了一定的理论分析. 然而, 对于中间层、构造图等知识的解释十分有限, 背后的蒸馏机制尚不清楚. 因此, 蒸馏效果背后的数学原理在很大程度上未被充分探索, 图知识蒸馏方法的理论研究仍然值得进一步探究和关注, 这对探索新的高效图蒸馏方法具有重要指导意义.

(5) 图蒸馏学习新范式. 由于图知识蒸馏在许多任务中表现出了令人印象深刻的性能改善, 大量的研究人员开始尝试将其与现有的深度学习新技术相结合, 包括对抗学习 (adversarial learning)、神经架构搜索 (neural architecture search)、图神经网络 (graph network networks)、强化学习 (reinforcement learning)、增量学习 (incremental learning)、联邦学习 (federated learning)、量化与剪枝 (quantization and prunning) 等. 知识蒸馏技术与其他技术结合使用, 衍生出了大量具有实用价值的应用. 例如, 知识蒸馏可以作为一种有效策略防御深度神经网络中的对抗扰动[127,128], 并可以用于解决数据隐私和安全问题[129,130]. 但是这些方法目前还在探索阶段, 方法还不成熟. 因此, 如何将知识蒸馏与其他技术方案更好地结合, 对于图蒸馏扩展到其他用途和应用是一个很有价值和意义的未来方向.

9 总 结

本文从图数据和知识蒸馏的基本概念出发, 对图知识蒸馏方法进行了全面的梳理分析. 首先, 根据图蒸馏算法的设计特点, 可以将其划分为面向深度神经网络的图知识蒸馏、面向图神经网络的图知识蒸馏和基于图知识的模型自蒸馏这 3 大类方法. 其次, 根据方法对知识蒸馏位置的处理手段, 可进一步细分为输出层、中间层和构造图知识方法. 接着, 通过实验对比了主流图知识蒸馏方法的算法性能. 此外, 还总结了图知识蒸馏在其他领域的重要应用场景. 最后, 对近年来图知识蒸馏学习的研究方向进行了总结和展望. 希望本文可以给图表示学习和知识蒸馏的研究人员提供一些参考, 促进该领域的持续发展.

References:

[1] Aggarwal CC, Wang HX. Managing and Mining Graph Data. New York: Springer, 2010. [doi: 10.1007/978-1-4419-6045-0]

[2] Fan WQ, Ma Y, Li Q, He Y, Zhao E, Tang JL, Yin DW. Graph neural networks for social recommendation. In: Proc. of the 2019 World Wide Web Conf. San Francisco, ACM, 2019. 417–426. [doi: 10.1145/3308558.3313488]

[3] Zhao TY, Hu Y, Valsdottir LR, Zang TY, Peng JJ. Identifying drug–target interactions based on graph convolutional network and deep neural network. Briefings in Bioinformatics, 2021, 22(2): 2141–2150. [doi: 10.1093/bib/bbaa044]

[4] Cui ZY, Henrickson K, Ke RM, Wang YH. Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. IEEE Trans. on Intelligent Transportation Systems, 2020, 21(11): 4883–4894. [doi: 10.1109/TITS.2019.2950416]

[5] Shi WJ, Rajkumar R. Point-GNN: Graph neural network for 3D object detection in a point cloud. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1708–1716. [doi: 10.1109/CVPR42600.2020.00178]

[6] Mirhoseini A, Goldie A, Yazgan M, Jiang JW, Songhori E, Wang S, Lee YJ, Johnson E, Pathak O, Nazi A, Pak J, Tong A, Srinivasa K,

Hang W, Tuncer E, Le QV, Laudon J, Ho R, Carpenter R, Dean J. A graph placement methodology for fast chip design. Nature, 2021, 594(7862): 207–212. [doi: 10.1038/s41586-021-03544-w]

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

[8] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans. on Neural Networks, 2009, 20(1): 61–80. [doi: 10.1109/TNN.2008.2005605]

[9] Oono K, Suzuki T. Graph neural networks exponentially lose expressive power for node classification. arXiv:1905.10947, 2021.

[10] Zhang MH, Chen YX. Link prediction based on graph neural networks. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 5171–5181.

[11] Errica F, Podda M, Bacciu D, Micheli A. A fair comparison of graph neural networks for graph classification. arXiv:1912.09893, 2022.

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

[13] Chen GB, Choi W, Yu X, Han T, Chandraker M. Learning efficient object detection models with knowledge distillation. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 742–751.

[14] Chebotar Y, Waters A. Distilling knowledge from ensembles of neural networks for speech recognition. In: Proc. of the 2016 Interspeech. San Francisco, 2016. 3439–3443. [doi: 10.21437/Interspeech.2016-1190]

[15] Liu XD, He PC, Chen WZ, Gao JF. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. arXiv:1904.09482, 2019.

[16] Yang YD, Qiu JY, Song ML, Tao DC, Wang XC. Distilling knowledge from graph convolutional networks. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7072–7081. [doi: 10.1109/CVPR42600.2020.00710]

[17] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907, 2017.

[18] Fan SH, Zhu JX, Han XT, Shi C, Hu LM, Ma BY, Li YL. Metapath-guided heterogeneous graph neural network for intent recommendation. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019. 2478–2486. [doi: 10.1145/3292500.3330673]

[19] Wang F, Yang JF, Wang MY, Jia CY, Shi XX, Hao GF, Yang GF. Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction. Science Bulletin, 2020, 65(14): 1184–1191. [doi: 10.1016/j.scib.2020.04.006]

[20] Battaglia P, Pascanu R, Lai M, Rezende DJ, Kavukcuoglu K. Interaction networks for learning about objects, relations and physics. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 4509–4517.

[21] Lin X, Quan Z, Wang ZJ, Ma TF, Zeng XX. KGNN: Knowledge graph neural network for drug-drug interaction prediction. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. 2020. 2739–2745. [doi: 10.24963/ijcai.2020/380]

[22] Zhang G, He H, Katabi D. Circuit-GNN: Graph neural networks for distributed circuit design. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 7364–7373.

[23] Estrach JB, Zaremba W, Szlam A, LeCun Y. Spectral networks and deep locally connected networks on graphs. In: Proc. of the 2nd Int'l Conf. on Learning Representations. 2014.

[24] Chung FRK. Spectral Graph Theory. Providence: American Mathematical Society, 1997. https://www.google.com.au/books/edition/Spectral_Graph_Theory/4IK8DgAAQBAJ

[25] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016. 3844–3852.

[26] Li RY, Wang S, Zhu FY, Huang JZ. Adaptive graph convolutional neural networks. Proc. of the AAAI Conf. on Artificial Intelligence, 2018, 32(1): 3546–3553. [doi: 10.1609/aaai.v32i1.11691]

[27] Zhuang CY, Ma Q. Dual graph convolutional networks for graph-based semi-supervised classification. In: Proc. of the 2018 World Wide Web Conf. Lyon: Int'l World Wide Web Conf. Steering Committee, 2018. 499–508. [doi: 10.1145/3178876.3186116]

[28] Xu BB, Shen HW, Cao Q, Qiu YQ, Cheng XQ. Graph wavelet neural network. arXiv:1904.07785, 2019.

[29] Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 1025–1035.

[30] Chen J, Ma TF, Xiao C. FastGCN: Fast learning with graph convolutional networks via importance sampling. arXiv:1801.10247, 2018.

[31] Zou DF, Hu ZN, Wang YW, Jiang S, Sun YZ, Gu QQ. Layer-dependent importance sampling for training deep and large graph convolutional networks. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems, 2019. 11249–11259.

[32] Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. arXiv:1710.10903, 2018.

[33] Atwood J, Towsley D. Diffusion-convolutional neural networks. In: Proc. of the 30th Int'l Conf. on Neural Information Processing

Systems. Barcelona: Curran Associates Inc., 2016. 2001–2009.

[34] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 1263–1272.

[35] Wang X L, Girshick R, Gupta A, He KM. Non-local neural networks. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803. [doi: 10.1109/CVPR.2018.00813]

[36] Battaglia PW, Hamrick JB, Bapst V, *et al*. Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261, 2018.

[37] Zhou J, Cui GQ, Hu SD, Zhang ZY, Yang C, Liu ZY, Wang LF, Li CC, Sun MS. Graph neural networks: A review of methods and applications. AI Open, 2020, 1: 57–81. [doi: 10.1016/j.aiopen.2021.01.001]

[38] Wu ZH, Pan SR, Chen FW, Long GD, Zhang CQ, Yu PS. A comprehensive survey on graph neural networks. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(1): 4–24. [doi: 10.1109/TNNLS.2020.2978386]

[39] Zhang ZW, Cui P, Zhu WW. Deep learning on graphs: A survey. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(1): 249–270. [doi: 10.1109/TKDE.2020.2981333]

[40] Skarding J, Gabrys B, Musial K. Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey. IEEE Access, 2021, 9: 79143–79168. [doi: 10.1109/ACCESS.2021.3082932]

[41] Yang C, Xiao YX, Zhang Y, Sun YZ, Han JW. Heterogeneous network representation learning: A unified framework with survey and benchmark. IEEE Trans. on Knowledge and Data Engineering, 2022, 34(10): 4854–4873. [doi: 10.1109/TKDE.2020.3045924]

[42] Sun LC, Dou YT, Yang C, Wang J, Liu YX, Yu PS, He LF, Li B. Adversarial attack and defense on graph data: A survey. arXiv:1812.10528, 2022.

[43] Wu LF, Chen Y, Shen K, Guo XJ, Gao HN, Li SC, Pei J, Long B. Graph neural networks for natural language processing: A survey. arXiv:2106.06090, 2022.

[44] Nazir U, Wang H, Taj M. Survey of image based graph neural networks. arXiv:2106.06307, 2021.

[45] Lopera DS, Servadei L, Kiprit GN, Hazra S, Wille R, Ecker W. A survey of graph neural networks for electronic design automation. In: Proc. of the 3rd ACM/IEEE Workshop on Machine Learning for CAD (MLCAD). Raleigh: IEEE, 2021. 1–6. [doi: 10.1109/MLCAD52597.2021.9531070]

[46] Wu SW, Sun F, Zhang WT, Xie X, Cui B. Graph neural networks in recommender systems: A survey. ACM Computing Surveys, 2022, 55(5): 97. [doi: 10.1145/3535101]

[47] Lamb LC, d'Avila Garcez A, Gori M, Prates MOR, Avelar PHC, Vardi MY. Graph neural networks meet neural-symbolic computing: A survey and perspective. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. 2020. 4877–4884. [doi: 10.24963/ijcai.2020/679]

[48] Yuan L, Tay FEH, Li GL, Wang T, Feng JS. Revisiting knowledge distillation via label smoothing regularization. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3902–3910. [doi: 10.1109/CVPR42600.2020.00396]

[49] Zhang Y, Xiang T, Hospedales TM, Lu HC. Deep mutual learning. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4320–4328. [doi: 10.1109/CVPR.2018.00454]

[50] Furlanello T, Lipton ZC, Tschannen M, Itti L, Anandkumar A. Born again neural networks. arXiv:1805.04770, 2018.

[51] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. arXiv:1412.6550, 2015.

[52] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv:1612.03928, 2017.

[53] Kim J, Park SU, Kwak N. Paraphrasing complex network: Network compression via factor transfer. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 2765–2774.

[54] Passban P, Wu YM, Rezagholizadeh M, Liu Q. ALP-KD: Attention-based layer projection for knowledge distillation. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(15): 13657–13665. [doi: 10.1609/aaai.v35i15.17610]

[55] Yim J, Joo D, Bae J, Kim J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7130–7138. [doi: 10.1109/CVPR.2017.754]

[56] Park W, Kim D, Lu Y, Cho M. Relational knowledge distillation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3962–3971. [doi: 10.1109/CVPR.2019.00409]

[57] Liu YF, Cao JJ, Li B, Yuan CF, Hu WM, Li YX, Duan YQ. Knowledge distillation via instance relationship graph. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7089–7097. [doi: 10.1109/CVPR.2019.00726]

[58] Tung F, Mori G. Similarity-preserving knowledge distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1365–1374. [doi: 10.1109/ICCV.2019.00145]

[59] Peng BY, Jin X, Li DS, Zhou SF, Wu YC, Liu JH, Zhang ZN, Liu Y. Correlation congruence for knowledge distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 5006–5015. [doi: 10.1109/ICCV.2019.00511]

[60] Passalis N, Tzelepi M, Tefas A. Probabilistic knowledge transfer for lightweight deep representation learning. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(5): 2030–2039. [doi: 10.1109/TNNLS.2020.2995884]

[61] Chen HT, Wang YH, Xu C, Xu C, Tao DC. Learning student networks via feature embedding. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(1): 25–35. [doi: 10.1109/TNNLS.2020.2970494]

[62] Zhang ZY, Shu XB, Yu BW, Liu TW, Zhao JP, Li QG, Guo L. Distilling knowledge from well-informed soft labels for neural relation extraction. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(5): 9620–9627. [doi: 10.1609/aaai.v34i05.6509]

[63] Minami S, Hirakawa T, Yamashita T, Fujiyoshi H. Knowledge transfer graph for deep collaborative learning. In: Proc. of the 15th Asian Conf. on Computer Vision. Kyoto: Springer, 2020. 203–217. [doi: 10.1007/978-3-030-69538-5_13]

[64] Wang HY, Lian DF, Ge Y. Binarized collaborative filtering with distilling graph convolutional networks. In: Proc. of the 28th Int'l Joint Conf. on Artificial Intelligence. Macao: AAAI Press, 2019. 4802–4808.

[65] Pan BX, Cai HY, Huang DA, Lee KH, Gaidon A, Adeli E, Niebles JC. Spatio-temporal graph for video captioning with knowledge distillation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10867–10876. [doi: 10.1109/CVPR42600.2020.01088]

[66] Koji T, Kanji T. Dark reciprocal-rank: Teacher-to-student knowledge transfer from self-localization model to graph-convolutional neural network. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation (ICRA). Xi'an: IEEE, 2021. 1846–1853. [doi: 10.1109/ICRA48506.2021.9561158]

[67] Lee S, Song BC. Interpretable embedding procedure knowledge transfer via stacked principal component analysis and graph neural network. Proc. of the AAAI Conf. on Artificial Intelligence, 2021, 35(9): 8297–8305. [doi: 10.1609/aaai.v35i9.17009]

[68] Zhou S, Wang YC, Chen DF, Chen JW, Wang X, Wang C, Bu JJ. Distilling holistic knowledge with graph neural networks. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 10367–10376. [doi: 10.1109/ICCV48922.2021.01022]

[69] Lee S, Song BC. Graph-based knowledge distillation by multi-head attention network. arXiv:1907.02226, 2019.

[70] Chen YX, Chen PG, Liu S, Wang LW, Jia JY. Deep structured instance graph for distilling object detectors. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 4339–4348. [doi: 10.1109/ICCV48922.2021.00432]

[71] Passalis N, Tzelepi M, Tefas A. Heterogeneous knowledge distillation using information flow modeling. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2336–2345. [doi: 10.1109/CVPR42600.2020.00241]

[72] Xue MQ, Song J, Wang XC, Chen Y, Wang X, Song ML. KDExplainer: A task-oriented attention model for explaining knowledge distillation. arXiv:2105.04181, 2021.

[73] Song J, Zhang HF, Wang XC, Xue MQ, Chen Y, Sun L, Tao DC, Song ML. Tree-like decision distillation. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 13483–13492. [doi: 10.1109/CVPR46437.2021.01328]

[74] Zhu YS, Zhang W, Chen MY, Chen H, Cheng X, Zhang W, Chen HJ. DualDE: Dually distilling knowledge graph embedding for faster and cheaper reasoning. In: Proc. of the 15th ACM Int'l Conf. on Web Search and Data Mining. Virtual Event: ACM, 2022. 1516–1524. [doi: 10.1145/3488560.3498437]

[75] Guo D, Wang H, Wang M. Context-aware graph inference with knowledge distillation for visual dialog. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6056–6073. [doi: 10.1109/TPAMI.2021.3085755]

[76] Lassance C, Bontonou M, Hacene GB, Gripon V, Tang J, Ortega A. Deep geometric knowledge distillation with graphs. In: Proc. of the 2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 8484–8488. [doi: 10.1109/ICASSP40776.2020.9053986]

[77] Chen MY, Zhang W, Zhu YS, Zhou HT, Yuan ZG, Xu CL, Chen HJ. Meta-knowledge transfer for inductive knowledge graph embedding. In: Proc. of the 45th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Madrid: ACM, 2022. 927–937. [doi: 10.1145/3477495.3531757]

[78] Zhang CR, Peng YX. Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. 2018. 1135–1141. [doi: 10.24963/ijcai.2018/158]

[79] He TY, Chen JL, Tan X, Qin T. Language graph distillation for low-resource machine translation. arXiv:1908.06258, 2019.

[80] Luo ZL, Hsieh JT, Jiang L, Niebles JC, Fei-Fei L. Graph distillation for action detection with privileged modalities. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 174–192. [doi: 10.1007/978-3-030-01264-9_11]

[81] Liu XB, Zhang SL. Graph consistency based mean-teaching for unsupervised domain adaptive person re-identification. arXiv:2105.

04776, 2021.

[82] Zhang YF, Jiang M, Zhao Q. Saliency prediction with external knowledge. In: Proc. of the 2021 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2021. 484–493. [doi: 10.1109/WACV48630.2021.00053]

[83] Hou YN, Ma Z, Liu CX, Hui TW, Loy CC. Inter-region affinity distillation for road marking segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12483−12492. [doi: 10.1109/CVPR42600.2020.01250]

[84] Tu K, Cui P, Wang DX, Zhang ZQ, Zhou J, Qi Y, Zhu WW. Conditional graph attention networks for distilling and refining knowledge graphs in recommendation. In: Proc. of the 30th ACM Int'l Conf. on Information & Knowledge Management. Virtual Event: ACM, 2021. 1834–1843. [doi: 10.1145/3459637.3482331]

[85] Deng X, Zhang ZF. Graph-free knowledge distillation for graph neural networks. arXiv:2105.07519, 2021.

[86] Zhang WT, Miao XP, Shao YX, Jiang JW, Chen L, Ruas O, Cui B. Reliable data distillation on graph convolutional network. In: Proc. of the 2020 ACM SIGMOD Int'l Conf. on Management of Data. Portland: ACM, 2020. 1399–1414. [doi: 10.1145/3318464.3389706]

[87] Ghorbani M, Bahrami M, Kazi A, Baghshah MS, Rabiee HR, Navab N. GKD: Semi-supervised graph knowledge distillation for graph-independent inference. In: Proc. of the 24th Int'l Conf. on Medical Image Computing and Computer-assisted Intervention. Strasbourg: Springer, 2021. 709–718. [doi: 10.1007/978-3-030-87240-3_68]

[88] Zhang SC, Liu Y, Sun YZ, Shah N. Graph-less neural networks: Teaching old MLPs new tricks via distillation. arXiv:2110.08727, 2022.

[89] Antaris S, Rafailidis D. Distill2Vec: Dynamic graph representation learning with knowledge distillation. In: Proc. of the 2020 IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM). The Hague: IEEE, 2020. 60–64. [doi: 10.1109/ASONAM49781.2020.9381315]

[90] Zhan K, Niu CX. Mutual teaching for graph convolutional networks. Future Generation Computer Systems, 2021, 115: 837–843. [doi: 10.1016/j.future.2020.10.016]

[91] Yan BC, Wang CK, Guo GY, Luo YK. TinyGNN: Learning efficient graph neural networks. In: Proc. of the 26th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Virtual Event: ACM, 2020. 1848–1856. [doi: 10.1145/3394486.3403236]

[92] Ma RR, Pang GS, Chen L, van den Hengel A. Deep graph-level anomaly detection by glocal knowledge distillation. In: Proc. of the 15th ACM Int'l Conf. on Web Search and Data Mining. Virtual Event: ACM, 2022. 704–714. [doi: 10.1145/3488560.3498473]

[93] Zhang CH, He YF, Cen YK, Hou ZY, Feng WZ, Dong YX, Cheng X, Cai HY, He F, Tang J. SCR: Training graph neural networks with consistency regularization. arXiv:2112.04319, 2022.

[94] Zhang WT, Jiang YZH, Li Y, Sheng ZA, Shen Y, Miao XP, Wang L, Yang Z, Cui B. ROD: Reception-aware online distillation for sparse graphs. In: Proc. of the 27th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining. Virtual Event: ACM, 2021. 2232–2242. [doi: 10.1145/3447548.3467221]

[95] Li Y, Liu L, Wang GY, Du Y, Chen PG. EGNN: Constructing explainable graph neural networks via knowledge distillation. Knowledge-based Systems, 2022, 241: 108345. [doi: 10.1016/j.knosys.2022.108345]

[96] Wang YN, Zhang YX, Coates M. Graph structure aware contrastive knowledge distillation for incremental learning in recommender systems. In: Proc. of the 30th ACM Int'l Conf. on Information & Knowledge Management. Virtual Event: ACM, 2021. 3518−3522. [doi: 10.1145/3459637.3482117]

[97] Kim J, Jung J, Kang U. Compressing deep graph convolution network with multi-staged knowledge distillation. PLoS ONE, 2021, 16(8): e0256187. [doi: 10.1371/JOURNAL.PONE.0256187]

[98] Antaris S, Rafailidis D, Girdzijauskas S. EGAD: Evolving graph representation learning with self-attention and knowledge distillation for live video streaming events. In: Proc. of the 2020 IEEE Int'l Conf. on Big Data (Big Data). Atlanta: IEEE, 2020. 1455–1464. [doi: 10.1109/BigData50022.2020.9378219]

[99] Jing YC, Yang YD, Wang XC, Song ML, Tao DC. Amalgamating knowledge from heterogeneous graph neural networks. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15704−15713. [doi: 10.1109/CVPR46437.2021.01545]

[100] Zheng WQ, Huang EW, Rao N, Katariya S, Wang ZY, Subbian K. Cold brew: Distilling graph node representations with incomplete or missing neighborhoods. arXiv:2111.04840, 2022.

[101] Wang S, Zhang K, Wu L, Ma HP, Hong RC, Wang M. Privileged graph distillation for cold start recommendation. In: Proc. of the 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Virtual Event: ACM, 2021. 1187–1196. [doi: 10.1145/3404835.3462929]

[102] Wang C, Wang Z, Chen DF, Zhou S, Feng Y, Chen C. Online adversarial distillation for graph neural networks. arXiv:2112.13966,

2021.

[103] Wang C, Zhou S, Yu K, Chen DF, Li BL, Feng Y, Chen C. Collaborative knowledge distillation for heterogeneous information network embedding. In: Proc. of the 2022 ACM Web Conf. ACM, 2022. 1631–1639. [doi: 10.1145/3485447.3512209]

[104] Bahri M, Bahl G, Zafeiriou S. Binary graph neural networks. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 9487–9496. [doi: 10.1109/CVPR46437.2021.00937]

[105] Qin C, Zhao HD, Wang LC, Wang H, Zhang YL, Fu Y. Slow learning and fast inference: Efficient graph similarity computation via knowledge distillation. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021. 14110–14121.

[106] Huang ZH, Tang YH, Chen YW. A graph neural network-based node classification model on class-imbalanced graph data. Knowledge-based Systems, 2022, 244: 108538. [doi: 10.1016/j.knosys.2022.108538]

[107] Ma JQ, Mei QZ. Graph representation learning via multi-task knowledge distillation. In: Proc. of the 33rd Conf. on Neural Information Processing Systems. 2019.

[108] Yao HX, Zhang CX, Wei Y, Jiang M, Wang SH, Huang JZ, Chawla N, Li ZH. Graph few-shot learning via knowledge transfer. Proc. of the AAAI Conf. on Artificial Intelligence, 2020, 34(4): 6656–6663. [doi: 10.1609/aaai.v34i04.6142]

[109] Wang JJ, Wang XF, Jin B, Yan JC, Zhang WJ, Zha HY. Heterogeneous graph-based knowledge transfer for generalized zero-shot learning. In: Proc. of the 25th Int'l Conf. on Pattern Recognition (ICPR). Milan: IEEE, 2021. 1859–1866. [doi: 10.1109/ICPR48806.2021.9412524]

[110] Yang C, Liu JW, Shi C. Extract the knowledge of graph neural networks and go beyond it: An effective knowledge distillation framework. In: Proc. of the 2021 Web Conf. Ljubljana: ACM, 2021. 1227–1237. [doi: 10.1145/3442381.3450068]

[111] Song QQ, Su J, Zhang W. scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. Nature Communications, 2021, 12(1): 3826. [doi: 10.1038/S41467-021-24172-Y]

[112] Qian YY, Zhang YM, Ye YF, Zhang CX. Distilling meta knowledge on heterogeneous graph for illicit drug trafficker detection on social media. In: Proc. of the 2021 Advances in Neural Information Processing Systems. 2021. 26911–26923.

[113] Joshi CK, Liu FY, Xun X, Lin J, Foo CS. On representation knowledge distillation for graph neural networks. arXiv:2111.04964, 2023.

[114] Liu J, Zheng TY, Hao QF. HIRE: Distilling high-order relational knowledge from heterogeneous graph neural networks. Neurocomputing, 2022, 507: 67–83. [doi: 10.1016/j.neucom.2022.08.022]

[115] Luo Y, Chen AG, Yan K, Tian L. Distilling self-knowledge from contrastive links to classify graph nodes without passing messages. arXiv:2106.08541, 2021.

[116] Zhang HL, Lin S, Liu WY, Zhou P, Tang J, Liang XD, Xing EP. Iterative graph self-distillation. arXiv:2010.12609, 2023.

[117] Yu L, Pei SC, Ding LZ, Zhou J, Li LF, Zhang CX, Zhang XL. SAIL: Self-augmented graph contrastive learning. Proc. of the AAAI Conf. on Artificial Intelligence, 2022, 36(8): 8927–8935. [doi: 10.1609/aaai.v36i8.20875]

[118] Chen YZ, Bian YT, Xiao X, Rong Y, Xu TY, Huang JZ. On self-distilling graph neural network. In: Proc. of the 30th Int'l Joint Conf. on Artificial Intelligence. 2021. 2278–2284.

[119] Ren YT, Ji JZ, Niu LF, Lei ML. Multi-task self-distillation for graph-based semi-supervised learning. arXiv:2112.01174, 2022.

[120] Krizhevsky A. Learning multiple layers of features from tiny images. Technical Report, Toronto: University of Toronto, 2009.

[121] Yang ZL, Cohen WW, Salakhudinov R. Revisiting semi-supervised learning with graph embeddings. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 40–48.

[122] Shchur O, Mumme M, Bojchevski A, Günnemann S. Pitfalls of graph neural network evaluation. arXiv:1811.05868, 2019.

[123] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[124] Li MY, Lin J, Ding YY, Liu ZJ, Zhu JY, Han S. GAN compression: Efficient architectures for interactive conditional GANs. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5283–5293. [doi: 10.1109/CVPR42600.2020.00533]

[125] Cheng X, Rao ZF, Chen YL, Zhang QS. Explaining knowledge distillation by quantifying the knowledge. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 12922–12932. [doi: 10.1109/CVPR42600.2020.01294]

[126] Mobahi H, Farajtabar M, Bartlett PL. Self-distillation amplifies regularization in Hilbert space. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 3351–3361.

[127] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proc. of the 2016 IEEE Symp. on Security and Privacy (SP). San Jose: IEEE, 2016. 582–597. [doi: 10.1109/SP.2016.41]

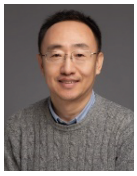
[128] Ross A, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input

gradients. Proc. of the AAAI Conf. on Artificial Intelligence, 2018, 32(1): 1660–1669. [doi: 10.1609/aaai.v32i1.11504]

[129] Papernot N, Abadi M, Erlingsson Ú, Goodfellow I, Talwar K. Semi-supervised knowledge transfer for deep learning from private training data. arXiv:1610.05755, 2017.

[130] Wang J, Bao WD, Sun LC, Zhu XM, Cao BK, Yu PS. Private model compression via knowledge distillation. Proc. of the AAAI Conf. on Artificial Intelligence, 2019, 33(1): 1190–1197. [doi: 10.1609/aaai.v33i01.33011190]

刘静(1994－), 女, 博士生, CCF 学生会员, 主要研究领域为图神经网络, 异构图表示学习, 知识蒸馏.



郝沁汾(1969－), 男, 博士, 研究员, CCF 高级会员, 主要研究领域为计算机体系结构, 图计算.

郑铜亚(1997－), 男, 博士, CCF 学生会员, 主要研究领域为图神经网络, 时序图, 可解释人工智能.