基于图的知识蒸馏：调查与实验评价

1. 介绍

对基于图的知识蒸馏方法进行了全面的概述，并进行系统地分类和总结，同时讨论了它们的局限性和未来的方向；首先介绍了图神经网络和知识蒸馏的背景; 全面总结了3类基于图的知识蒸馏方法，即基于图的深度神经网络知识蒸馏(Graph-based Knowledge Distillation for deep neural networks, DKD)、基于图的GNNs知识蒸馏(Graph-based Knowledge Distillation for GNNs, GKD)和基于自知识蒸馏的图知识蒸馏(Self-Knowledge Distillation)；每一类方法根据输出层、中间层和构建的图进一步分为知识蒸馏方法。随后，对各种基于图的知识蒸馏算法的思想进行了分析比较，并通过实验结果分析了各算法的优缺点。此外，还列举了基于图的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，对基于图的知识蒸馏的发展进行总结和展望。

2. 背景

2.1 图神经网络

- 用于处理非欧氏空间的非结构化数据
- 通常被表示为结点和边的集合，常用于图分析
- **应用领域**：电子商务推荐——需要一个基于图形的学习系统，通过利用用户和项目之间的交互来实现高度精确的推荐、生物化学、物理建模、知识图谱、电路设计
- **挑战**：图大小不同、节点无序、每个节点的邻居节点个数不同——GNN
- GNN方法：**光谱法**(the spectral method)和**空间法**(the spatial methods)
- **光谱法**：
    - 基于光谱的方法从图信号处理的角度引入滤波器来定义图的卷积；
    - 2013年，Bruna等[86]基于谱理论[87]将频域卷积运算的概念引入gnn中——首次提出了第一种光谱方法光谱CNN。
    - ChebyNet [88]利用切比雪夫多项式的矩阵形式进行参数化核卷积，大大降低了谱CNN的参数和计算复杂度，从而使谱方法实用
    - 缺点：谱方法在计算时通常需要同时对整个图进行处理，并且需要承受矩阵分解的高时间复杂度，这很难并行或扩展到大图。
- **空间法**：
    - 基于空间的方法直接对图进行卷积操作，并将图的卷积表示为聚合来自邻域的特征信息。
    - GCN [17]利用**一阶逼近**进一步简化了谱域内的图卷积，使图卷积操作可以在空间域内进行，大大提高了图卷积模型的计算效率。
    - 加速训练方法

      抽样策略：将计算限制在一批节点而不是整个图上来实现高效的计算（缓解诸如训练时间和内存需求等问题）——SAGE [92], FastGCN [93], LADIES [94]
    - 基于空间的方法具有较高的自由度、良好的可计算性和较高的推理效率
- 现有GNN模型的局限性：
    - 现有的GNN大多是半监督学习，使其性能很大程度上依赖于高质量的标注数据
    - 随着图尺度的发展，现有图模型的设计越来越复杂，给图模型的计算和图存储带来了一定的挑战。

2.2 知识蒸馏

- 一种模型压缩方法，知识蒸馏（KD）使用T-S框架对一个大型教师模型进行预训练，以提炼得到一个轻量级的学生模型，提高了学生模型的泛化能力，获得了更好的性能和更高的精度。通过蒸馏，将教师模型中的"知识"（软标签监督信息）转移到学生模型中；得到一个复杂度较低的学生模型
- 根据不同的知识转移方式分为两条技术路线：**基于响应的蒸馏**和**基于特征的蒸馏方法**
- **基于响应的蒸馏**：
    - **标签平滑**(label smoothing)——使用教师模型的输出概率（软标签）作为平滑标签来训练学生模型
        - 标签平滑（Label smoothing），像L1、L2和dropout一样，是机器学习领域的一种**正则化方法**，通常用于分类问题，目的是防止模型在训练时过于自信地预测标签，改善泛化能力差的问题。
        - 将0，1分布修改成概率分布的形式；标签平滑后的分布就相当于往真实分布中加入了噪声，使得预测正负样本的输出值差别不那么大，从而避免过拟合，提高模型的泛化能力。
        - [12]是Hinton于2015年提出的知识蒸馏的开创性工作，首次提出将教师模型的softmax层的输出概率转移到学生模型作为"软化"，以提高学生模型的性能。
    - 从学生模型中学习反馈信息：
        - DML [112]提出了深度互学习的策略：允许**一组学生**在网络上同时进行训练，通过监督真实标签和同伴网络输出结果的学习经验，实现相互学习和进步。
        - BAN [113]采用集成的方法对学生模型进行训练，使其网络结构与教师模型相同，在计算机视觉和语言建模下游任务中明显优于教师模型。
- **基于特征的蒸馏方法**
    - 教师网络结构中**中间层特征表示**中包含的语义信息作为知识传递到学生模型。
        - 目前已成为主流方法，包括注意力机制、概率分布匹配
        - FitNet [114]是第一个采用该方法的经典工作，利用教师网络的输出和中间层的特征嵌入作为监督信息来扩展KD，实现了深度模型网络压缩问题。
        - 从基于特征的蒸馏方法中得到了新的**基于关系的蒸馏方法**[18-22,118,119]，但它们都是为了更好地将基于特征的知识从教师那里提炼给学生。

3. 基于图的知识蒸馏

随着知识蒸馏技术的发展，仅从单个样品中提取信息的蒸馏方法不再适用，因为它们提供的信息有限；为了提取不同数据样本之间丰富的相关信息，提出了基于关系的知识蒸馏方法[18-22,118,119]，该方法通过隐式/显式构建样本之间的关系图，充分挖掘教师网络中样本之间的结构特征知识。

基于深度神经网络（DNNs）的关系式知识蒸馏方法和基于GNNs的知识蒸馏方法称为基于图的知识蒸馏方法。

基于图的知识蒸馏法旨在将教师模型中直接/间接构建的**样本关系**语义信息提取到学生模型中，以获得更普遍、更丰富、更充分的知识。

- GNN依赖于高质量的标签数据和复杂的网络模型；存在标签难以获取和计算资源成本高昂的问题
- 面对GNN中稀疏数据标签和模型计算的高复杂性问题，如何设计性能保证的更小更快的网络已成为研究的热点。
- 基于图的知识蒸馏分类：
    - 基于图的深度神经网络知识蒸馏（DKD）
        - 基于标签/响应的蒸馏

Hinton在Distilling the Knowledge in a Neural Network中提出知识蒸馏，旨在将隐藏在大型网络（教师模型，t）中的知识转移到轻量级网络（学生模型，s）中，从而使学生模型获得更好的性能。通过一个系数τ来软化教师模型的输出概率，τ越大，输出概率越平滑。

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

通过加权教师模型中的软目标和真实目标损失作为学生模型的损失函数进行训练能使学生模型表现效果更好

$$\mathcal{L}_{KD}=L_{CE}\left(p_s,y\right)+\alpha\tau^2 KL\left(p_s^{\tau},p_t^{\tau}\right),$$

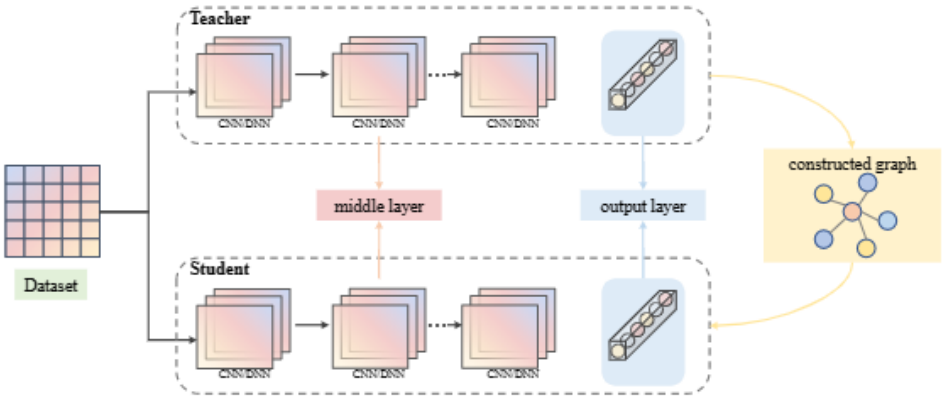
α是调整两个损失函数比率的超参数。KL是Kullback-Leibler散度。

局限性：侧重单样本学习，无法学习样本之间的关系

- 基于特征的蒸馏和基于关系的蒸馏（隐式构造图方法）

  使学生模型模拟教师模型中样本之间的相似性，而不是模拟教师模型的单个样本的输出。借助于隐式/显式样本关系构建图，学生模型可以充分挖掘教师网络中样本之间的结构化特征信息，并实现从教师模型中提取的通用、丰富和充分的知识，以指导学生模型。

  - 首先，基于教师和学生模型分别在CNN/DNN框架下获得的样本特征表示，构建了各自的样本关系图，不同颜色的顶点表示不同的训练样本。
  - 其次，使用相似度函数分别计算教师和学生网络样本之间的相似度。
  - 最后，使用距离测量函数最小化学生和教师的特征分布，以确保学生模型能够学习教师模型特征空间中多个样本的相关性。



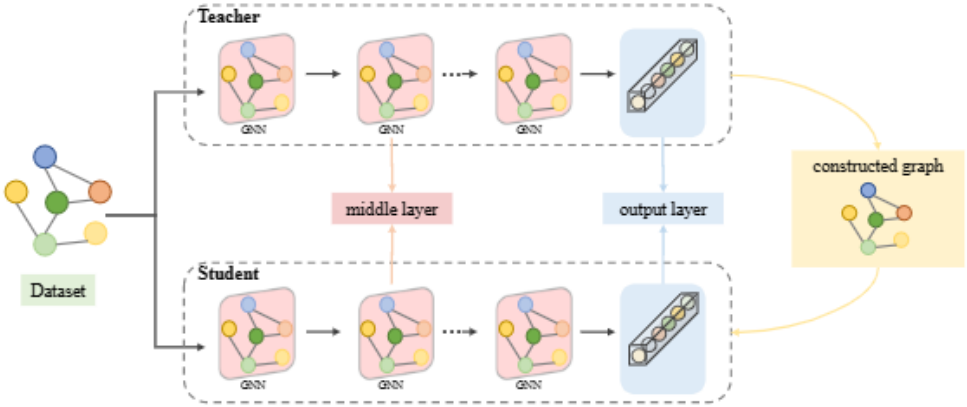
  - 大多数隐式/显式构造的图方法发生在中间卷积层。学生模型可以通过使用构建的图来直接提取由教师模型学习的丰富的样本间相关性知识，而不是仅仅拟合教师模型中单个样本的输出概率分布。学生模型可以捕获教师模型的输入样本之间的空间几何知识，以更准确地测量样本特征之间的相似度，并提高学生模型的知识蒸馏学习效果。
  - 挑战：如何正确和恰当地构造辅助图来建模关系数据的结构知识

- 基于图的神经网络知识蒸馏（GKD）
  - 图表示学习遵循消息传递范式：通过自身节点特征和邻接节点特征的多层聚合来更新下一层节点

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

GNN缺点：GNN的性能严重依赖于大量高质量的标记数据和高度复杂的网络模型

- 可以直接对图形数据进行建模，以自然地挖掘教师模型中的图形结构知识信息，并将其传递给学生模型，与DKD相似，GKD从图卷积的中间层和输出层提取信息构建一个图结构
  - 首先，基于具有GNN框架的教师和学生模型的中间特征表示，构建其各自的节点间关系图（如图3所示），其中不同的颜色表示不同的节点（不同类型的异构节点在异构图中表示）
  - 其次，采用相似性函数来度量师生网络内部拓扑节点之间的相关性
  - 使用距离测量函数来计算教师和学生各自的内部节点嵌入之间的差异
  - 累加用于转移知识层的所有损失，并将拓扑知识和节点关系知识迁移到学生模型中。



  - 损失函数：

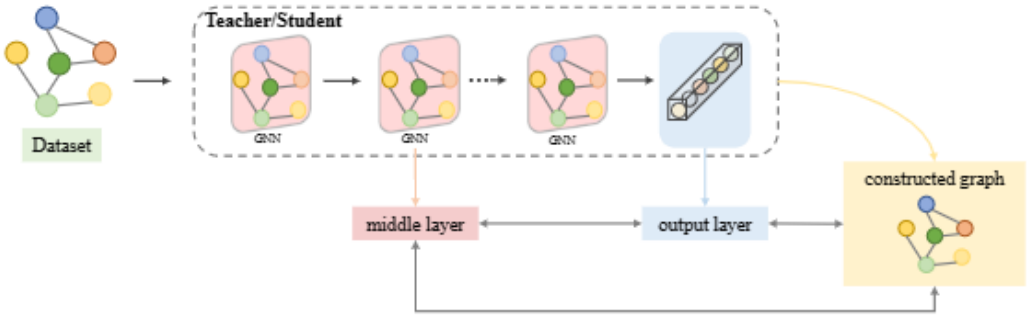
$$\mathcal{L}_G = \sum_{l \in L} \sum_{(x,x') \in x^2} D_G \left(S\left(x_s^l, x'^l_s\right), S\left(x_t^l, x'^l_t\right) \right).$$

  - GNN是一个强大的图形建模工具，它可以直接在中间/输出层进行提取，并将图节点之间的拓扑知识传递给学生模型。为了进一步探索特征空间中局部节点之间的关系，人们在构造中间卷积层节点之间的关联图以提取它们之间的相关知识方面做了大量的工作，如LSP[16]、HIRE[75]等。
  - 挑战：如何在GNN上充分挖掘图拓扑和语义信息以进行知识转移

- 基于图的自知识蒸馏（SKD）

  自知识蒸馏是基于T-S架构的基于图的知识蒸馏方法的一种特殊情况，它是指在没有额外教师模型的帮助下进行知识转移的特殊蒸馏方法。自知识蒸馏意味着单个网络模型既是学生模型，也是教师模型。它通常在自己的深层和浅层之间传递信息，以指导自己的学习，而无需教师模型的帮助。自知识蒸馏法简单高效，已成为当前实际落地项目的首选。

  - 基于GNN框架下的中间层/输出层特征表示，构建节点间关系图，其中不同的颜色表示不同的节点（异构图中表示不同类型的异构节点）
  - 利用相似性函数来测量特征空间中GNN的浅层和深层的内部拓扑节点表示之间的相似性。
  - 通过使用距离度量函数来计算浅网络和深网络之间的差异，并且可以通过多次迭代计算来学习更多样的知识。



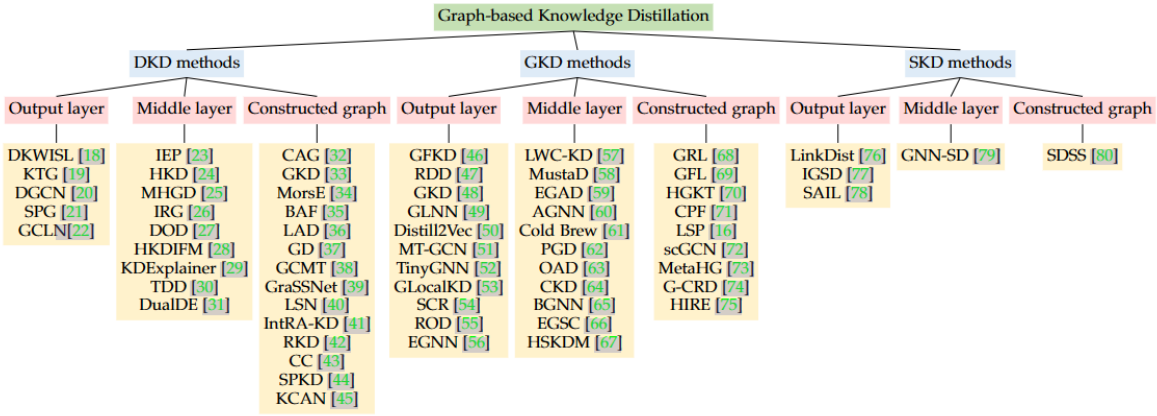
- 损失函数

$$\mathcal{L}_{Self} = \sum_{l=1,\ldots,l-1} D_{Self}\left(G_s^l(X), G_t^{l+1}(X)\right).$$

- 优点：与传统的两阶段T-S蒸馏模式相比，自知识蒸馏学习模式可以大大节省模型训练时间，大大提高训练效率，并在没有老师指导的情况下实现模型性能的提高

  缺点：缺乏丰富的外部知识，如果可以明确地引入外部知识，例如与知识图谱相结合，可能有助于提高学生模型的性能；传统的两阶段T-S蒸馏模式和自知识蒸馏模式的优缺点尚无定论，并且缺乏在相同实验环境和任务下对它们进行比较分析，值得进一步研究；自知识蒸馏的可解释性分析需要进一步研究

4. 方法



4.1 基于图的深度神经网络知识提取

根据知识蒸馏的位置，将基于图的深度神经网络知识蒸馏（DKD）分为输出层、中间层和构造的图知识。

- 三种类型知识转移形式：输出层、中间层和构造图
- 相同点：对于基于输出层、中间层和构造图的DKD方法，每种类型的基于图的知识提取算法都基于相同位置的知识提取。它们都遵循DKD模型框架和图蒸馏损失L_D范式。

  不同点：对于每种类型的DKD，它们的差异反映在许多方面，例如具体实现、距离度量函数、下游任务、应用程序等。

  - 输出层知识
    - DKWISL[18]通过使用KL距离度量将KD应用于NLP中的关系提取
    - KTG[19]通过使用KL散度测量教师和学生之间的分布差异，用于协作学习的图像识别应用
    - GCLN[22]使用欧氏距离将KD应用于视觉机器人定位场景，以进行图像语义分割

- 中间层知识
  - IEP[23]结合KL和L1，将知识应用于多任务学习中的转移学习和图像分类
  - HKD[24]利用InfoCE在图形推理的视觉对话任务中引入知识蒸馏技术
  - IRG[26]提出Hit将知识蒸馏应用于图像识别场景
- 构建图
  - CAG[32]提出了构建的基于图形的知识蒸馏技术，以增强下游图形推理任务中学生模型的视觉对话性能
  - GKD[33]探索Frobenius以最小化教师和学生之间的分布差异，并压缩学生模型
  - MorsE[34]使用L2来传递元知识，以改进链接预测和问答系统任务中的学生模型。

Method	Distillation Location			Distance Measurement	Task	Application
	Output Layer	Intermediate Layer	Constructed Graph			
IEP [23]		✓		KL, L1	Multi-task learning	Transfer learning, image classification
HKD [24]		✓		InfoCE	Knowledge distillation	Image classification, knowledge transfer
CAG [32]			✓	KL	Graph inference	Visual dialogue
DKWISL [18]	✓			KL	Natural language processing	Relation extraction
KTG [19]	✓			KL	Collaborative learning	Image recognition
MHGD [25]		✓		KL	Multi-task learning	Image recognition
IRG [26]	✓	✓		Hit	Knowledge distillation	Image recognition
DGCN [28]	✓			KL	Collaborative filtering	Item recommendations
GKD [33]	✓	✓	✓	Frobenius	Model compression	Image classification
SPG [21]	✓			KL	Natural language processing	Video captioning
MorsE [34]			✓	L_2	Meta-knowledge transfer	Link prediction, question answering system
GCLN [22]	✓			L_2	Image semantic segmentation	Vision robot self-positioning
DOD [27]	✓	✓		KL	Object detection	Object Detectors
BAF [35]			✓	EMD	Model compression	Video classification
LAD [36]			✓	BELU	Natural language processing	Machine translation
GD [37]			✓	Cosine	Multimodal video	Motion detection, action classification
GCMT [38]	✓	✓	✓	CE	Unsupervised domain adaptation	Person re-identification
GraSSNet [39]			✓	MSE	Knowledge transfer	Saliency prediction
LSN [40]	✓	✓	✓	KL, MSE	Model compression	Node classification
IntRA-KD [41]		✓	✓	MSE	Model compression	Road marking segmentation
RKD [42]	✓		✓	Euclidean,Huber	Knowledge distillation	Image classification, few-Shot Learning
CC [43]	✓	✓	✓	KL, MSE	Knowledge distillation	Image classification, person re-identification
SPKD [44]			✓	Frobenius	Knowledge distillation	Image classification, transfer learning
HKDIFM [28]		✓		KL	Knowledge distillation	Image classification
KDExplainer [29]	✓	✓		CE, KL	Interpretability	Image classification
TDD [30]	✓	✓		CE, KL	Interpretability	Image classification
DualDE [31]		✓		JSD	Knowledge distillation	Node classification, link prediction
KCAN [45]			✓	BPR	Knowledge graph	Top-K Recommendation, TR Prediction

4.2 基于图的图神经网络知识提取

对于GNN中的知识蒸馏方法，仍然根据知识蒸馏的位置将其分类为基于输出层知识的、基于中间层知识的和基于构造图的。

- 相同点：对于基于输出层、中间层和构造图的GKD方法，每种基于图的知识提取算法都基于在相同位置提取知识。

  不同点：不同的方法使用的度量函数和具体实现（知识转移形式不同）

Method	Distillation Location			Distance Measurement	Task	Application
	Output Layer	Intermediate Layer	Constructed Graph			
GFKD [46]	✓			KL	Data-free distillation	Zero-Shot learning
LWC-KD [57]		✓		MSE, L_2	Incremental learning	Recommender system
MustaD [58]	✓	✓		KL	Model compression	Node classification
RDD [47]	✓			MSE	Semi-supervised learning	Node classification
EGAD [59]		✓		RMSE, MAE	Semi-supervised learning	Live video streaming events
GRL [68]			✓	MAE	Multi-task learning	Graph-level prediction
GFL [69]			✓	Frobenius	Few-Shot learning	Node classification
HGKT [70]			✓	Wasserstein	Zero-Shot learning	Node classification
AGNN [60]	✓	✓		MSE	Model compression	Node classification, point cloud classification
CPF [71]	✓	✓	✓	L_2, KL	Knowledge distillation	Node classification
LSP [16]		✓	✓	KL	Model compression	Node classification, point cloud classification
GKD [48]	✓			CE	Graph inference	Disease diagnosis and prediction
scGCN [72]			✓	CE	Single cell omics	Cell identification, cross-species classification
MetaHG [73]			✓	KL	Illegal drug trafficker	Classification
Cold Brew [61]	✓	✓		MSE, CE	Cold start	Recommender system
PGD [62]		✓		MSE	Cold start	Recommender system
GLNN [49]	✓			KL	Offline knowledge distillation	Node classification
Distill2Vec [50]	✓			KL	Model compression	Link prediction
MT-GCN [51]	✓			KL	Semi-supervised learning	Node classification
TinyGNN [52]	✓			CE	Model compression	Node classification
GLocalKD [53]	✓			KL	Anomaly detection	Anomaly detection
OAD [63]	✓	✓		CE, KL	Online adversarial distillation	Node classification
SCR [54]	✓			MSE	Model training	Node classification
ROD [55]	✓			KL	Model compression	Node classification, clustering, link prediction
EGNN [56]	✓			KL	Model interpretability	Node classification
CKD [64]		✓		JSD	Knowledge distillation	Node classification, link prediction
G-CRD [74]	✓	✓	✓	InfoCE	Model compression	Classification, similarity measures
BGNN [65]		✓		KL	Model compression	Image classification
EGSC [66]		✓		MSE, Huber	Model compression	Anomaly detection, graph similarity calculation
HSKDM [67]		✓		KL, Triplet	Knowledge distillation	Node classification
HIRE [75]	✓	✓	✓	KL, MSE	Knowledge distillation	Node classification, clustering, visualization

4.3 基于图的自知识蒸馏

基于图的自知识蒸馏SKD方法分为三种类型：输出层、中间层和根据蒸馏位置构建的图

- 相同点：对于基于输出层、中间层和构造图的SKD方法，每种基于图的知识提取算法都基于在相同位置提取知识。

  不同点：

- SAIL[78]、GNN-SD[79]和SDSS[80]分别对输出层、中间层和构建的图知识使用KL散度。
- LinkDist[76]使用MSE距离应用GKD进行节点分类和模型压缩。
- IGSD[77]使用MSE来缓解过度平滑。GNN-SD[79]分别使用KL和L2距离度量进行图分类和分子性质预测。
- SDSS[80]在下游半监督学习中利用KL和MSE距离度量进行多任务节点分类

Method	Distillation Location			Distance Measurement	Task	Application
	Output Layer	Intermediate Layer	Constructed Graph			
LinkDist [76]	✓			MSE	Model compression	Node classification
IGSD [77]	✓			InfoCE	Graph-level task	Graph classification, molecular property prediction
GNN-SD [79]		✓		KL, L_2	Relieve over-smoothing	Node & graph classification
SDSS [80]	✓	✓	✓	KL, MSE	Semi-supervised learning	Multitask node classification
SAIL [78]	✓			KL	Unsupervised learning	Node classification & clustering, link prediction

5. 实验

分别比较和分析了深度神经网络的基于图的知识蒸馏方法（DKD）、基于图神经网络的知识蒸馏（GKD）和基于自知识蒸馏的基于图知识蒸馏（SKD）在相关数据集上的表现

5.1 数据集

为了比较使用DKD前后的实验效果，选择了DNN中常用的两个数据集，包括CIFAR-10[120]和CIFAR-100[120]。

Dataset	$\|Total\|$	$\|Train\|$	$\|Test\|$	Class
Cora	60000	50000	10000	10
Citeseer	60000	50000	10000	100

为了比较GKD方法的实验结果，GNN中常用的七个数据集，包括Cora、Citeseer、Pubmed、AmazonPhoto（A-P）、AmazonComputers（A-C）、（Physics）and（CS）

Dataset	$\|V\|$	$\|E\|$	Feature	Class
Cora	2708	5278	1443	7
Citeseer	3327	4552	3703	6
Pubmed	19717	44324	500	3
A-P	7650	119043	745	8
A-C	13752	245778	767	10
Physics	34493	247962	8415	5
CS	18333	81894	6805	15

Cora：是一个由机器学习论文组成的基准引用数据集[121]。节点表示论文，边缘表示引用关系。每个节点都有一个1433维的特征，并且类标签指示每个论文所属的研究领域。任务是根据引文网络将论文分类到不同的领域。

Citeseer：是另一个常用的基准引用数据集[121]。每个节点表示一篇论文，每个边缘表示两篇论文之间的引用关系，节点特征维度为3703，有六个类标签，任务是预测出版物的类别。

Pubmed：也是一个引用网络[121]，包含19717个节点和44324条边，其中节点表示糖尿病相关论文，边表示引用论文之间的关系。该节点的特征在于具有500维的TF/IDF加权词频。类别标签有三个类别

A-P和A-C：是亚马逊的产品采购网络[122]。节点表示商品，边缘表示两者经常一起购买。节点特征由产品评论的单词包表示，任务是预测项目的类别。

Physics and CS：是从2016年KDD杯挑战赛的Microsoft学术图表中提取的常用引用网络[122]，节点指示作者，边表示作者是否处于合作关系；节点特征由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，任务是将作者分成各自的研究领域。

5.2 实验设置与结果

DKD：选择代表性的Resnet-20[123]模型作为深度神经网络模型的框架，以测试基于图的知识蒸馏方法在两个数据集（CIFAR-10和CIFAR-100）的图像分类任务上的性能并使用精度度量。为了比较蒸馏效果，论文中选择了经典的KD和IRG、RKD和CC，这三种常用的DKD方法。

Model	Metric	CIFAR-10	CIFAR-100
Teacher	Accuracy	0.9237	0.6892
+KD	Accuracy	**0.9330**	0.7036
+IRG	Accuracy	0.9277	**0.7037**
+RKD	Accuracy	0.9272	0.6948
+CC	Accuracy	0.9301	0.6927

GKD：在节点分类任务的比较实验中，在下表所述的七个数据集上选择了GNN的最具代表性的模型（即，GCN[17]、GAT[95]和SAGE[92]），利用经典的KD和CPF蒸馏方法，选择F1微观和F1宏观指标来评估蒸馏效果。

Model	Metric	Cora	Citeseer	Pubmed	A-P	A-C	Physics	CS
Node clustering of graph distillation variants based on GCN								
GCN	NMI	0.5568	0.4291	0.3711	0.4235	0.3399	0.7029	0.6736
	ARI	0.5120	0.4241	0.4094	0.2619	0.2083	0.6806	0.5336
+KD	NMI	**0.6011**	0.4655	**0.3874**	**0.5932**	0.4578	0.7111	**0.7145**
	ARI	**0.5933**	0.4620	**0.4401**	**0.4655**	0.2883	0.6890	**0.6025**
+CPF	NMI	0.5988	**0.4714**	*0.1935*	0.5888	**0.5299**	**0.7519**	*0.5299*
	ARI	0.5801	***0.4721***	*0.1214*	0.3872	**0.2985**	***0.8228***	*0.2985*
Node clustering of graph distillation variants based on GAT								
GAT	NMI	0.6056	0.4297	0.3626	0.6545	0.4975	0.7669	0.7531
	ARI	0.5634	0.4257	0.3910	0.5311	0.4018	0.8391	0.6889
+KD	NMI	**0.6145**	0.4550	0.3754	**0.6814**	**0.5567**	**0.7711**	**0.7719**
	ARI	**0.5799**	**0.4449**	0.4169	**0.5975**	**0.4767**	**0.8506**	**0.7930**
+CPF	NMI	0.6066	**0.4551**	**0.4021**	*0.5113*	0.4981	*0.6147*	*0.5850*
	ARI	*0.5109*	*0.4177*	***0.4266***	*0.2884*	*0.2994*	*0.5654*	*0.4371*
Node clustering of graph distillation variants based on SAGE								
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*

SKD：选择经典的GCN模型作为图神经网络模型框架，并将节点分类作为任务，以测试对Cora、Citeseer和Pubmed数据集的蒸馏效果。选择精度作为分类度量。应用了经典的KD和LinkDist、SAIL和SDSS自知识蒸馏蒸馏方法。

Model	Metric	Cora	Citeseer	Pubmed
Teacher	Accuracy	0.8183	0.6762	0.7859
+KD	Accuracy	*0.8005*	0.6821	*0.7571*
+LinkDist	Accuracy	*0.7572*	0.7119	*0.7484*
+SAIL	Accuracy	<u>0.8463</u>	<u>0.7424</u>	**0.8381**
+SDSS	Accuracy	**0.8600**	**0.7613**	<u>0.8221</u>

6. 应用

Field	Problem	Backbone	Algorithm
Computer Vision	Image classification	CNN	HKD [24], GKD [33], SPKD [44], HKDIFM [28], KDExplainer [29], TDD [30]
	Image recognition	CNN	KTG [19], MHGD [25], IRG [26]
	Robot localization	GNN	GCLN [22]
	Object detection	CNN	DOD [27], GD [37]
	Video classification	CNN	BAF [35]
	Event prediction	GNN	EGAD [59]
	Person re-identification	CNN	GCMT [38], CC [43]
	Road marking	CNN	IntRA-KD [41]
Natural Language Processing	Visual dialogue	CNN	CAG [32]
	Relation extraction	CNN	DKWISL [18]
	Video captioning	CNN	SPG [21]
	Machine translation	CNN	LAD [36]
	Metric learning	CNN	RKD [42]
Recommender System	Incremental learning	GNN	LWC-KD [57]
	Collaborative filtering	CNN	DGCN [20]
	Cold start	GNN	PGD [62]
	Tail generalization	GNN	Cold Brew [61]
Multi-task Learning	Transfer learning	CNN	IEP [23]
	Image recognition	GNN	GRL [68]
	Image recognition	CNN	MHGD [25]
	Self knowledge distillation	GNN	SDSS [80]
Zero-Shot Learning	Data-free distillation	GNN	GFKD [46]
	Model enhancement	GNN	HGKT [70]

7. 未来与展望

- 蒸馏位置的确定

  通过对图蒸馏工作的归纳分析，现有的大多数图蒸馏方法使用不同类型的知识源组合，包括输出层、中间层和构建的图知识。然而，尚不清楚知识的哪个位置起着重要作用。具体选择哪一层进行蒸馏，目前很少有研究。如何设计一种更快、更通用、精度更可靠的图蒸馏方法，以同时对所有类型的知识源进行建模，仍然具有挑战性。

- 蒸馏模式的选择

  两种流行的图蒸馏方法是T-S蒸馏模式和自知识蒸馏模式。由于其灵活性、可控性和易用性，T-S蒸馏法适用于大型复杂的模型压缩，模型自知识蒸馏蒸馏方法由于其结构简单和有效的训练效率，在具有较大开销的下游业务场景中得到广泛应用。然而，这两种蒸馏方法仍存在不足：T-S训练复杂且耗时；自我知识蒸馏缺乏理论支持，仅限于教师和学生模型表现相当的问题场景。同时，目前缺乏对这两种蒸馏方法的比较研究。因此，有必要研究蒸馏模式的选择如何影响KD的有效性，以及如何设计有效的蒸馏框架。

- 蒸馏距离测量的选择

  损失函数的选择方法多种多样，如KL、MSE、InfoCE等，但对于在图蒸馏过程中应选择哪种损失函数以更好地指导学生的模型训练过程，仍然是个值得考虑的问题

8. 总结

本文基于图数据和知识蒸馏的基本概念，对基于图的知识蒸馏方法进行了深入的梳理。

根据基于图的知识蒸馏算法的设计特点，可以将其分为三类：基于图的深度神经网络知识蒸馏(Graph-based Knowledge Distillation for deep neural networks, DKD)、基于图的GNNs知识蒸馏(Graph-based Knowledge Distillation for GNNs, GKD)和基于自知识蒸馏的图知识蒸馏(Self-Knowledge Distillation，SKD)；并且进一步细分为输出层、中间层和基于知识蒸馏位置的构造图方法。然后，对主流的基于图的知识提取方法的算法性能进行了实验比较。并且总结了基于图形的知识蒸馏在各个领域的关键应用场景。最后，总结并展望了近年来基于图形的知识蒸馏学习的研究方向。