2.3 基于知识蒸馏的无监督图表示学习算法设计

* 无监督图表示学习任务

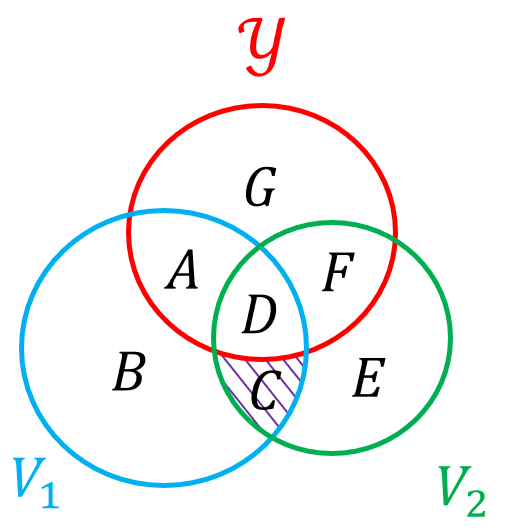
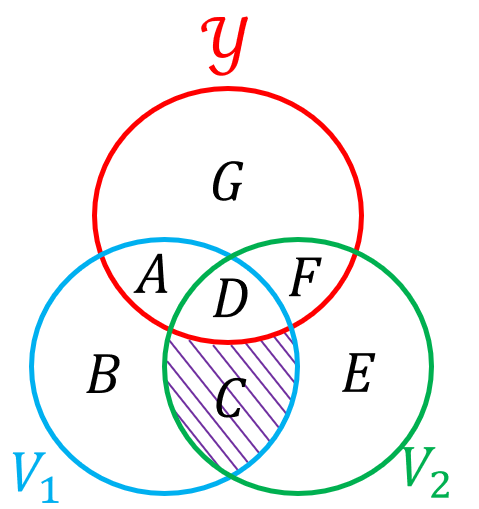
在2.2的研究内容中，研究了本课题在异质图半监督节点分类任务上的设计，但是异质图半监督节点分类属于异质图中特定的一个基准任务，在本课题的第二个研究内容中，我们将目光放到了更加普遍的任务-无监督图表示学习任务。该任务在给定图的情况下，无需手动准备标签（监督信号），就可以从图数据本身学习节点和图的表示，该表示可以继续用于节点分类，图分类等具体的下游基准任务。

形式化上，给定图，其中和分别是图中边和节点的集合，表示图的带有自环的邻接矩阵。每个节点都关联于一个特征向量,其中d是特征向量的维度。无监督图表示学习任务从给定的数据中学习每个节点的表示向量。

常见的无监督图表示学习算法大致分为生成式和对比式两种，生成式的代表方法是图自编码器，通过用重构输入数据，使得中包含输入数据中有用的信息。而对比式的代表方法则是图对比学习，通常，图对比学习会首先通过数据增强等各种方式生成两个对比的视图，然后最大化它们之间的互信息，使得中包含了两个不同视图中共有的有效信息，丢弃了两个视图中不重叠的无效信息（噪声）。本研究内容面向图对比学习算法，设计针对图对比学习的知识蒸馏师生模型和优化框架，并在各种下游基准任务中取得更好的效果。

具体来说，标准的图对比学习范式（如GraphCL，GRACE，GCA）通常分为两个视图，每一个视图通过将不同的图增强（对输入的图数据进行数据增强，比如删除边，掩盖部分节点特征等）输入到同一个编码器（图神经网络）中，得到两个不同视图下的节点表示，然后通过InfoNCE Loss（公式 8）近似的使两个视图的表示(和)之间的互信息最大化。

在对比学习范式中遵循的先验知识是，通过数据增强获得的两个视图中都包含对下游基准任务有效的信息，而两个视图中不重叠的那部分信息则是对任务无用的噪声。如图3（a），两个不同的视图和，其共享的信息是区域C和D，其中区域D中是对下游任务有用的信息，而不共享的信息是A，B，E，F，对比学习范式希望通过最大化两个视图共享的信息（CD）使得D的区域增加。但是显然，最大化区域CD并不意味着D增加，C可能也会增加，区域C代表着和虽然共享但对下游任务无用的的噪声信息，C的增加很显然会导致学到的表示中含有噪声。所以本研究内容的目标是根据以上描述的无监督图对比学习中的先验知识，设计合理有效的知识蒸馏框架和教师学生模型学得更好的（区域C小，区域D大）节点和图表示。



(a)对称 （b）非对称

图3.不同视图的信息文氏图

* 基于图模型增强的教师学生模型设计

相比于传统的图对比学习范式（如GraphCL，GRACE，GCA）把两个不同的数据增强输入同一个编码器模型，现在计算机视觉中已经有一些工作（如DINO，BYOL）结合了知识蒸馏和对比学习范式，使用了两个不同的编码器模型（Teacher和Student）。具体来说，这些工作将学生模型的指数移动平均值作为教师模型（梯度停止），并且用学生模型去预测教师模型的输出表示。但是在图表示学习领域还鲜有类似工作，受（DINO，BYOL）启发，我们在图表示学习领域，设计了一种基于图模型增强的教师-学生模型架构，通过三种不同并且很有效的模型增强策略，得到教师模型和学生模型，并且采用InfoNCE 损失函数最大化学生模型和教师模型输出的表示的互信息。具体来说，本研究内容采用传统的对比学习范式作为基准，采用图神经网络为编码器编码视图，通过三种模型增强策略（非对称，随机传播，随机排列）对同一个图神经网络做两次不同的模型增强得到 教师模型和学生模型，然后分别用教师和学生模型编码数据增强后的视图，得到的两个视图表示通过InfoNCE进行互信息最大化，训练后的编码器（不进行模型增强）编码得到的图和节点表示可以用于下游基准任务，流程示意图如图5所示。

Design and implementation of graph contrast learning algorithm based on model enhancement



图4: 基于知识蒸馏的无监督图表示学习算法流程图

由图4可知，教师模型和学生模型的设计基于原图神经网络模型的两种模型增强，接下来具体介绍本课题采用的三种模型增强算法以及它们的好处。

* 预备知识和符号定义：

已知图，其中和分别是图中边和节点的集合，表示图的带有自环的邻接矩阵。每个节点都关联于一个特征向量,其中d是特征向量的维度。

图神经网络(GNNs) GNNs将深度学习技术推广到图中进行节点编码。具体来说，GNN将多个传播层和变换层叠加在一起，然后应用到原始特征上，这里我们将传播层和变换层的算子分别记为和：

其中为节点嵌入，为图滤波器矩阵，为线性变换的权值矩阵，σ为非线性函数(如ReLU)。图滤波器是基于邻接矩阵A的常数矩阵，是可训练参数。的输出与输入矩阵的形状相同，的输出为维的矩阵。借助这两种算子，节点表示可以传播到相邻节点，并通过线性和非线性变换映射到新的嵌入空间。

现在我们可以将图神经网络形式化为多个和算子的函数组合。例如，一个 L层的图卷积网络 (GCN)可以形式化如下：

其中◦表示两个操作符的组合，表示GCN的第L层变换在GCN中，图滤波器矩阵将邻接矩阵A归一化为:，其中D为度矩阵，本课题中定义图滤波器为，其中π∈(0,1),是单位矩阵。

* 非对称模型增强策略：

策略：上文提到，如图4中所示，学习到的对比学习表示包括与任务相关的信息(区域D)和与任务无关的噪声(区域C)。直观上，这两个视图不能太远(D中的共享信息太少)，也不能太近(C的噪声太多)。具体来说：如果仅采用数据增强技术(如删除边或节点)用于生成视图，则不能在保持任务相关信息不变的情况下产生足够多的增强;并且如果采用同一个视图编码器对视图进行编码，会导致视图之间较为紧密，区域C（噪声）过大。为了解决这个问题，本课题使用共享参数的非对称视图编码器。如图4(b)所示，不同的视图编码器会使两个视图之间的距离变远，而共享的参数可以保证两个视图之间的距离不会太远。采用这种方式可以减轻噪声(C区域)。在具体实现中，本课题使用了两个GCN编码器，其中运算符数目相同，而运算符数目不同。

优势：为了简单起见，在没有非线性函数的情况下，算子中的权重矩阵可以分解为一个单独的权重矩阵。然后我们有两个编码器的嵌入, 其中, 是两个视图编码器中运算符的数量，并且。假设图滤波器可以通过特征向量分解分解为，其中是特征向量的酉矩阵，是降序排列的特征值对角矩阵。的特征值越大，对应于图拉普拉斯矩阵的特征值越小，这通常被认为是下游任务更重要的信息。在这里，我们的重点在于分析图信号，并假设一个独热节点特征向量。然后视图的嵌入可以写成, 。对比损失的目的是减小正例对之间的距离，根据这一点，我们可以重写对比损失函数如下：

其中表示矩阵的迹，条件方程,是为了避免平凡的全零解。则有以下定理：

定理1：在公式(11)中最优的如下：

其中是使得最小化的。

由以上定理，我们需要找到使最小化的，显而易见，方程偏好λ→0或λ→1，当我们采用图滤波器，则的特征值在[0,1]范围内下降。对于Cora和Citeseer等流行的图数据集，的特征值在实践中会落在的范围内。此外，我们还发现这些流行的图数据的滤波器有许多特征值接近1，例如，在Cora中的第100大的特征值仍然大于0.998。因此，可能是的前列。此时视图编码器生成的表示为：

此时生成的表示就可以完美过滤高频噪声。相反，如果采用的相同视图编码器，我们必须考虑图增广对表示带来的扰动。假设数据增强对特征值的影响为，则得到。因此，最优的是由滤波器和数据增强的扰动共同决定的，数据增强不可避免地将高频噪声纳入学习到的表示中。

* 随机传播策略：

策略：随机传播策略则是在每次的训练过程中，我们改变GNN编码器中传播层的数量。具体地说，我们在每次训练中随机抽样，作为的数量，而不是使用固定的传播层数量（两层GCN中为2）。本课题认为，在每次训练中改变传播层数可以丰富训练实例的多样性，从而有助于预测下游任务。

优势：是一个带有个传播算子的GNN编码器，它通过聚合节点的邻居来计算每个节点的节点表示，也就是说等价于作用于根节点

的深度为的局部计算树的函数，设为训练集。然后我们可以将中的元素从单个节点变换为节点对应的深度为的计算树，当操作符的数量在每次训练过程中从随机选择时，可以假设样本的集合每次训练都在变化。因此，训练集被放大了很多倍。事实上，如果进一步考虑数据增强(例如随即删除边和节点)，训练集将涉及到计算树及其多个子树。使用随机层数的训练集容量和使用固定层数的训练集容量之间可能会有更大的差距(甚至指数级)。

* 随机排列策略：

策略：本课题采用的随机排列策略建议使用具有不同排列的算子和的视图编码器用于生成图增广的表示 。形式上，如果视图编码器有个算子和个算子，那么可以写成：

我们对于两个视图编码器和采用不同的，直觉上，打乱传播和转换操作符的顺序不会改变输入图的语义，并且会干扰编码表示，可以看作是一种更安全的模型增强策略。

* 基于互信息最大化的优化框架设计

通过上文三种模型增强的策略，本课题将传统对比学习范式中的固定的编码器通过不同的模型增强变化成两个不同的模型（教师模型和学生模型），随后，采用InfoNCE 损失函数进行训练之后，使用未进行模型增强的编码器（GCN）作为节点表示的生成器，生成节点表示用于下游任务。