GraphRNN : Generating Realistic Graphs with Deep Auto-regressive Models

GraphRNN 使用深度自回归模型生成显示中的图

主线：

需求：对图上的复杂分布进行建模并从中采样

难点：图的非唯一性，高维性，给定图的边之间存在复杂的非局部依赖关系

解决：提出名为GraphRNN 的深度自回归模型（能以最小的结构假设逼近任何图的分布）

模型思路：通过对一组有代表性的图进行训练学习生成图，并将图的生成过程分解为一连串的节点和边的生成

评估性能：引入基准数据集，基线和基于最大平均差异的新的评估标准（衡量图集之间的距离）

优点：GraphRNN 显著优于所有基线，能学习生成符合目标集结构特征的各种图，同时可以扩展到比以前的深度模型大50倍的图

传统的图生成模型：无法直接从观测数据中学习生成模型。

VAE，GAN 在图像和文本数据等复杂领域的生成建模方面取得了重要的进展。

生成图的深度学习模型受限，（仅限于单个途中学习，或者包含节点个数有限），来源于主线的三个挑战。

GraphRNN 简介：一个用于学习生成模型的可扩展的框架。

GraphRNN 以自回归（或者递归）的方式对图进行建模，来捕捉图中所有节点和边的复杂联合概率。

GraphRNN 可以被看作一个分层模型，图级RNN保持图的状态并生成新的节点，边级RNN为每一个新生成的节点生成边。

GraphRNN 中引入了BFS 的节点排序，以提高可扩展性。BFS 方法通过将不同的表征折叠成唯一的BFS树（解决图的非唯一表征困难），BFS引起的树状结构允许使用者在训练期间限制每个节点的边预测数量。

GraphRNN 评估设定：通过比较两组图的度分布，聚类系数分布，基于最大平均差异（MMD）的变体的图案计数。（可以比较图统计分布的高阶矩，并提供简单比较平均值更严格的评价）。

符号和问题定义

无向图 G=（V，E）；节点集V 和 边集E。

使用邻接矩阵来表示图，（需要一个节点排序Π将节点映射到邻接矩阵的行或列）

可知一共可能有n! 个可能的节点排列组合的集合。

生成模型的目标：

基于一组从数据分布p(G)中抽样的观察图G0=（G1 … Gs）,在图上学习一个分布pmodel(G)，其中每个图Gi 可能有不同的节点数和边数。当G属于G0时，进一步假设可以以相同的概率观察到任何节点排序Π。因此生成模型需要能够生成图，每个图都可能有指数级的表示，这与以前的图像，文本，时间序列的生成模型不同。可以应用于单个或多个输入训练图。

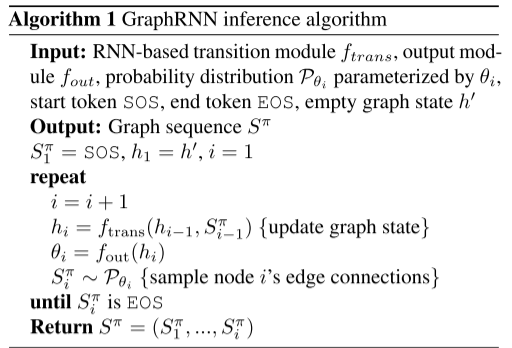
关键思想：

将不同的节点排序下的图表示成序列，然后在这些序列上建立一个自回归生成模型。

（1）将图建模成序列，论文中引入S^Π，其实就是原邻接矩阵的右上角，（无向图关于斜对角线对称），表示的是第i个节点和之前i-1 个节点的连接关系。据此，将一个图出现的概率p(G)表示为联合概率分布p(G,S^Π)，再根据上方假设，简化表示后，当前节点新的边的概率取决于该节点已经有的边的情况。（使用RNN）

（2）GraphRNN 的框架（如何生成一个图？）

在（1）中将问题从生成图转换成生成边，但边的概率依旧复杂，此时作者提出使用神经网络对复杂分布建模进行参数化。



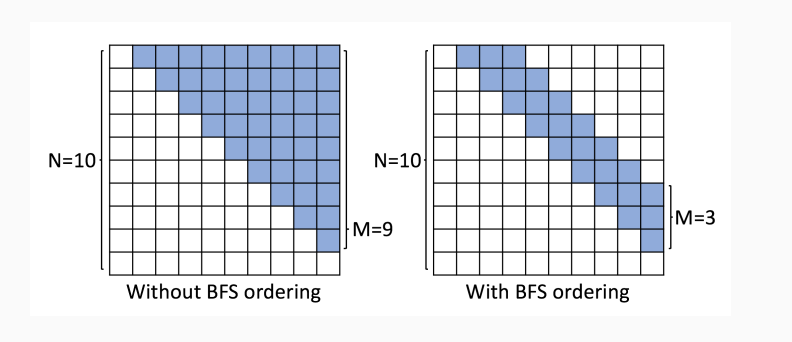
Ftrans 和 Fout 可以使用任意神经网络，Pi可以是二进制向量上的任意分布。

值得注意的点：论文中Ftrans,Fout使用RNN ，但是，RNN需要固定大小的输入向量，而之前定义的S^Π有不同的尺寸（右上角，多则i个，少则一个），使用引入BFS解决

（3）BFS 实现可操作性

训练生成模型，并不是学习在任何可能的节点序列下生成图，而是通过学习使用广度优先搜索（BFS）的节点序列来生成图。

优点：不局限于生成某类特定图集的图，更好的通用性。减少需要考虑的序列的总数量。BFS减少了边集RNN上对变得预测的数量来使学习更简单。同时，将大小不一的S重新定义为一个固定的M维向量，代表着节点Π（vi）与当前队列中BFS最大大小为M的节点之间的连接性。



GraphRNN 生成图的过程：

基于RNN 神经网络，从一个起始状态，利用RNN（graph-level RNN）逐步地生成图的状态，每一步向图中加入一个节点，再利用RNN（edge-level RNN）生成新的节点与已存在节点的连接关系，接着再向图中添加节点，重复下去一直到EOS。

结合上面的定义，hi 就是生成过程中图的状态，它是对已经生成的图的一个encode之后的结果，每一步输出的hi都要在进入另外一个RNN中作为输入，生成边的连接关系。

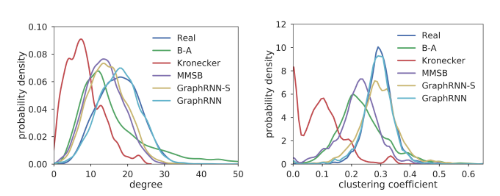
训练过程：（论文中没有详细的描述，需要翻看github代码）

首先将传统格式的图数据转换成论文中的序列形成的图，对于每个图，会对其节点进行多次排列，得到多次节点的不同序列，然后根据BFS算法进行遍历，可以得到不同排列顺序下的图生成序列。（将图概率问题转换成边缘分布问题），在训练中使用Teacher Forcing的技术进行训练。

训练结果：

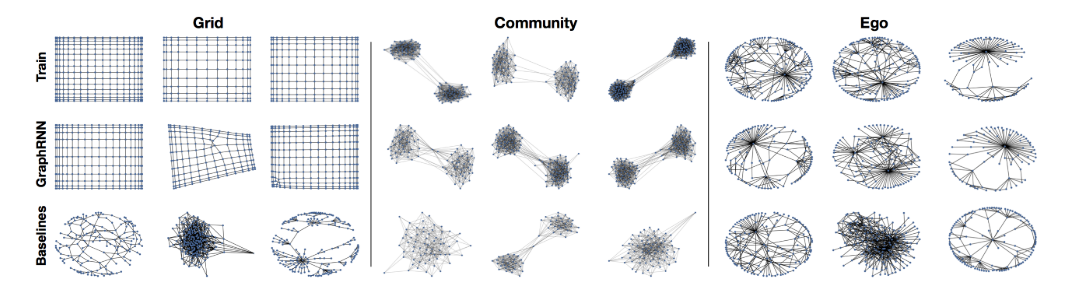
分为合成数据集和真实数据集，大小分布|V| { 10 - 2025 }；

评估方法：提出新的评价标准，比较其经验分布的所有时刻。基于最大平均差异（MMD）



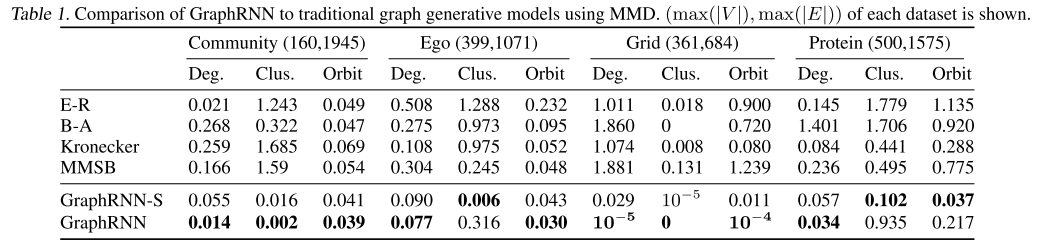
此为导出的MMD分数的度数和聚类系数分布，以及平均轨道计数统计（所有4个节点的轨道的出现次数）

表明GraphRNN 即使在数百个不同大小的图中，仍然可以很好的学习捕获基本的图统计量，生成的平均统计量与整个测试集的分布密切匹配。



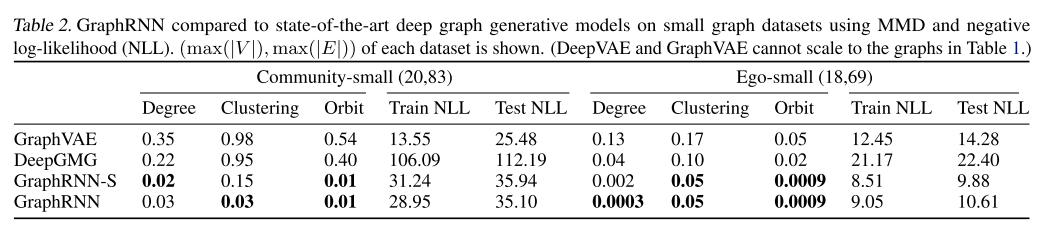
表明GraphRNN 能够捕捉到具有巨大差异的数据集的结构，有效学习像网格这样的常规结构和Ego这样更自然地结构。

（从第一个可以看出学习对未见过的网格宽度/高度进行归纳）

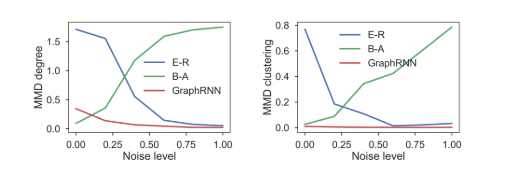


GraphRNN 和传统的生成模型比较，取得了最好的性能，与传统基线相比，MMD下降80%，与深度学习基线相比，MMD下降90%。

在Protein(蛋白)中（生物表示学方面的研究），GraphRNN-S效果更好，可能是蛋白质的数据集近似于欧氏空间，不涉及高度的复杂的边缘依赖关系。（个别的模型在特定的数据集上表现好，但不一定其他类型上通用）



比较训练集和测试集的负对数似然（NLLs），GraphRNN可以很好的泛化，平均NLL差距小22%。



随机扰动有100个节点的B-A图的（10%，20%，，，，100%），引入噪声后，GraphRNN 的稳定性保持强大的性能，表明了高鲁棒性和通用性。

总结：

论文提出了GraphRNN，一个针对图结构数据的自回归生成模型，以及一系列新的针对图生成问题的综合评估。

实验证明GraphRNN 比以前的先进模型相比取得了明显的性能，同时具有可扩展性和对噪声的鲁棒性。

思考：

论文是否解决了一开始的三个难题（图的非唯一性，高维性，给定图的边之间存在复杂的非局部依赖关系）：

1.高维性：（RNN？对图进行编码）（疑问）

2.非唯一性表示：（使用序列表示图的生成过程）

3.复杂的依赖关系：（使用BFS ，固定值做超参数）

未来工作方向：

生成更大规模的图，高效的生成指定条件的图。