实验五

实验要求

在给定的三个图数据下，使用GCN模型做节点分类与链路预测；

分析自环、层数、DropEdge、PairNorm、激活函数等因素对GCN在两个下游任务性能的影响。

Cora数据集

由机器学习论文组成，是近年来图深度学习很喜欢使用的数据集。该数据集共2708个样本点，每个样本点都是一篇科学论文，所有样本点被分为8个类别，类别分别是：
1）基于案例；2）遗传算法；3）神经网络；4）概率方法；5）强化学习；6）规则学习；7）理论

每篇论文都由一个1433维的词向量表示，所以，每个样本点具有1433个特征。词向量的每个元素都对应一个词，且该元素只有0或1两个取值。取0表示该元素对应的词不在论文中，取1表示在论文中。所有的词来源于一个具有1433个词的字典。

每篇论文都至少引用了一篇其他论文，或者被其他论文引用，也就是样本点之间存在联系，没有任何一个样本点与其他样本点完全没联系。如果将样本点看做图中的点，则这是一个连通的图，不存在孤立点。

CiteSeer数据集

引文网络。论文分为六类：Agents、AI（人工智能）、DB（数据库）、IR（信息检索）、ML（机器语言）和HCI。
共包含3312篇论文，记录了论文之间引用或被引用信息。去除停用词和在文档中出现频率小于10次的词，整理得到3703个唯一词。

PPI(生物化学结构)

PPI网络是蛋白质相互作用（Protein-Protein Interaction,PPI）网络的简称。
PPI是指两种或以上的蛋白质结合的过程，通常旨在执行其生化功能。一般地，如果两个蛋白质共同参与一个生命过程或者协同完成某一功能，都被看作这两个蛋白质之间存在相互作用。多个蛋白质之间的复杂的相互作用关系可以用PPI网络来描述。
PPI数据集共24张图，每张图对应不同的人体组织，平均每张图有2371个节点，共56944个节点818716条边，每个节点特征长度为50，其中包含位置基因集，基序集和免疫学特征。基因本体基作为label(总共121个)。

（ppi数据集需要大家自己再进行一下数据结构的处理 可以仿照Cora和CiteSeer数据集）

注：数据集维度对不上可能是数据清理的方式不同，但一般情况下，GCN在cora数据集上做节点分类可以得到80%以上的准确率，在citeseer上做节点分类可以达到70%以上的准确率。如果准确率相差很大，可能是模型实现有误。

实验结果提交

- 实验报告(.pdf)：需要包含实验过程，数据预处理部分，分析各个因素对模型性能的影响
- 实验代码(.py)
- 以上结果打包为压缩包，发至dl2021_bigdata@163.com
- 压缩包命名格式请按：学号_姓名_实验五，截止日期为12.27晚23:59