# 实验四

## 实验要求

使用pytorch 或者 tensorflow 的相关神经网络库编写图卷积神经网络模型 GCN， 并在相应的图结构数据集上完成节点分类和链路预测任务，最后分析自环、层数、 DropEdge 、PairNorm 、激活函数等因素对模型的分类和预测性能的影响 。

## 实验步骤

1. **网络框架**：要求选择 pytorch 或 tensorflow 其中之一，依据官方网站的指引安装包。这个实验还需要安装 torch_geometric。（如果前面实验已经安装过，则这个可以跳过）
2. **数据准备：** 本次实验使用的数据包含三个常用的图结构数据集： Cora 、Citeseer 、PPI 。下面分别进行介绍。
   - **Cora：** 该数据集是由 2708 篇机器学习论文作为节点、论文间引用关系作为有向边构成的图数据 。 具体的数据 描述见https://relational.fit.cvut.cz/dataset/CORA 。 数据集下载链接 https://linqsdata.soe.ucsc.edu/public/lbc/cora.tgz 。 另外， 提供一个数据处理范例链接https://graphsandnetworks.com/thecoradataset/ 。请同学们仔细阅读相关材料，了解文件的具体结构和数据格式。
   - **Citeseer：** 该数据集是由3312篇论文及相互引用构成的图数据集。数据集下载链接 https://linqsdata.soe.ucsc.edu/public/lbc/citeseer.tgz 。文件的结构和数据格式与Cora类似。
   - **PPI：** PPI 网络是蛋白质相互作用（ Protein Protein Interaction,PPI ）网络的简称数描述可参考链接 https://blog.csdn.net/ziqingnian/article/details/112979175 。数据集下载链接 http://snap.stanford.edu/graphsage/ppi.zip 。
3. **数据预处理：** 你需要通过pytorch 或 tensorflow所提供的标准数据接口，将原始数据处理为方便模型训练脚本所使用的数据结构，如 torch.utils.data.Dataset 等。 由于这三个数据集是非常常见的公开数据集，你可以参考一些公开代码片段 ，尤其是 github 上典型的GCN 教程级实现或相关论文的源码 。
4. **图网络模型：** 搭建GCN模型，这一步可以参考网络上公开的源码。
5. **节点分类：** 在三个数据集上按照节点分类任务的需求自行划分训练集 、验证集、测试集，并用搭建好的GCN 模型进行节点分类。
6. **链路预测：** 在三个数据集上按照链路预测任务的需求自行划分训练集、验证集、测试集，并用搭建好的GCN 模型进行链路预测。
7. **测试性能：** 选择你认为最合适的（例如，在验证集上表现最好的）一组超参数，重新训练模型，并在测试集上测试（注意，这理应是你的实验中 唯一一次在测试集上的测试），并记录测试的结果 。

## 实验提交

本次实验截止日期为 **6月 6 日 23:59:59**，需提交代码源文件及实验报告到邮箱：proton00@mail.ustc.edu.cn，具体要求如下：

1. 全部文件打包在一个压缩包内，压缩包命名为 学号- 姓名 - exp4.zip
2. 代码仅包含 .py 文件，请勿包含实验中间结果（例如中间保存的数据集等），如果有多个文件，放在 src/ 文件夹内。
3. 代码中提供一个可以直接运行的并输出结果的 **main.py，结果包括训练集损失、验证集损失随 epoch 改变的曲线（保存下来）和测试集的评价指标。**
4. 代码中提供一个描述所有需依赖包的 requirements.txt，手动列入代码中用到的所有非标准库及版本或者使用 `pip freeze > requirements.txt` 命令生成。
5. 实验报告要求 pdf 格式，要求包含姓名、学号。内容包括简要的**实验过程**和**关键代码**展示，对超参数的**实验分析**，最优超参数下的训练集、验证集**损失曲线**以及测试集上的**实验结果。**

## 参考资料

往届同学的实验代码和报告： https://github.com/hehaha68/USTC_2022Spring_Introduction-to-Deep-Learning

提供的 Lab4_demo.ipynb

实验数据下载链接：https://rec.ustc.edu.cn/share/3bab12a0-ee13-11ed-b34a-1d166b75eb33