

## 深度学习导论 DS2001.01.2023SP——实验三

### 实验要求

使用 pytorch 或者 tensorflow 编写图卷积神经网络模型 GCN，并在图结构数据集 Cora 上完成**节点分类**和**链路预测**任务，研究自环、层数、DropEdge、PairNorm、激活函数等因素对**节点分类**和**链路预测**性能的影响。




- A. 节点分类：预测节点的类别或标签。
- B. 链路预测：预测两节点之间是否存在潜在的链接(边)。

### 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。本次实验推荐使用 GPU 版本完成，如果你没有 GPU 机器，可考虑使用一些云资源，例如 Google Colab。

2. **数据集**：本次实验的数据集为 Cora 数据集。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://lings-data.soe.ucsc.edu/public/lbc/cora.tgz>。解压后数据集文件夹如下：

 cora.cites	2007/2/19 6:34	CITES 文件
 cora.content	2007/2/19 6:34	CONTENT 文件
 README	2007/2/19 6:34	文件

其中，README 文件对其他两个文件的数据格式及数据标签做了介绍。

要求将数据集划分为**互不相交**的训练集、验证集和测试集。实验调参只能在**验证集**上完成。

提示：可直接使用 PyG 中封装好的 datasets 进行数据集的加载使用，以及 RandomLinkSplit 对数据集进行划分。

3. **GCN 模型搭建**：采用 pytorch 或 tensorflow 所封装的 module 编写模型，无需手动完成底层 forward、backward 过程，但要求**手动编写图卷积层**。

4. **模型训练**：选择你认为合适的损失函数进行训练。将生成的**训练集**输入搭建好的模型进行前向的 loss 计算和反向的梯度传播，从而训练模型，同时也建议使用网络框架封装的 optimizer 完成参数更新过程。

5. **调参分析**：将训练好的模型在**验证集**上进行测试，对自环、层数、DropEdge、PairNorm、激活函数等模型的超参数进行调整后，再重新训练、测试，并分析对模型性能的影响，节点分类任务评价指标选用**准确率（Accuracy）**，链路预测任务评价指标选用 **AUC**。

6. **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，此处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（**Loss&Acc/AUC**）。

## 实验提交

实验三截止时间：6月3日 23:59:59，**线下**完成代码检查（关键代码讲解+运行展示+结果展示），并需在 **bb** 系统提交源代码及实验报告，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-ex3.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）；如果有多个代码文件，放在 **src**/文件夹内。
3. 实验报告提交为.pdf 格式，包含学号、姓名，内容包括简要的实验过程、关键代码展示及**功能介绍、对超参数的调试分析**以及测试集上的实验结果。