前言

Tutorial 内容涵盖从 AI 入门到科研基础，带领你从零开始独立完成科研任务，全面体验完整的科研流程。实验室拥有多样化的研究方向，一定能找到符合你兴趣的领域。（近期组内的研究主力在于大模型时代下的图学习）现有的研究方向已在实验室内打下了坚实的基础，并配备丰富的计算资源。此外，老师和学长学姐们会随时提供答疑和指导。加入实验室没有年级和专业的限制，任何人都可以参与🤗🤗🤗 。

本科生科研主要培养目标，你将收获以下内容：

1. 了解人工智能基础知识，读懂前沿算法代码，学会运用前沿技术解决问题
2. 掌握科研基本流程，锻炼严谨的科研思维，熟练运用各类科研工具和文献技巧
3. 探索创新性 Idea 并实现，在人工智能顶级会议上发表论文，参与国内外学术会议交流学习
4. 转换科研项目成果（国家级大创），斩获各种竞赛奖项，大厂名组访问实习

本次培训时长大约为一个月，希望能在培训任务的指引下，带领你掌握以下内容：

- 深度学习/机器学习的基本原理：数据集，模型，训练原理等；重点掌握图机器学习方向的热点研究领域；
- 图机器学习：了解图机器学习的基本原理和当前热门的研究方向，能够实现简单、基本的图学习任务训练及测试 Pipeline（节点分类、链路预测等...）
- 代码基本功：对 Python、PyTorch、PyG 有基本的了解，清楚 Anaconda 的作用，能使用 Linux 系统的远程服务器进行深度学习任务；能在给出 GitHub 仓库的情况下，根据论文内容跑通代码；能对已有代码进行改进；
- 科研基础：了解什么是 arXiv、Overleaf、如何进行文献管理和阅读 (Zotero)、如何做文献笔记和论文调研报告；
- 基本的 PPT 制作技能：可以考虑使用 Powerpoint，或者 LaTeX 的 Beamer 包完成汇报 PPT 的制作

入门培训内容（主要考核是否对科研有兴趣，愿意投入）

1. 深度学习的基本原理和代码功底（两周左右）：
    - 参考李宏毅的机器学习课程，学习机器学习、深度学习的基本原理：只需学习 P1-5, P7-8, P12-16, P18-19, P30-31，其余课程目前不用学习。
    - 参考斯坦福CS224W，学习图机器学习的基本原理：只需学习 1. Introduction, 2. Node embeddings, 3. Graph neural networks, 4. A general perspective on GNNs, 6. Theory of GNNs, 14. Graph Transformers, 15. Scaling to large graphs，其余暂时不用学习。
    - 代码能力：
        - 搭建本地代码开发环境，强烈推荐使用 VSCode + Anaconda 对 Python 虚拟环境进行管理，可以参考配置教程。
        - 如果对Python语言没有太多了解，可以先学习Python基础教程，也可以学习https://cs231n.github.io/python-numpy-tutorial/和Jupyter Notebook教程。对Linux不熟练的，学习Linux基础教程。对Git不熟悉的，学习一小时Git教程。
        - PyTorch 是基于 Python 实现深度学习的最常用工具包，建议在掌握深度学习的基本原理，听(看)完了李宏毅老师(斯坦福大学 slides)课程的基础上，学习PyTorch官方文档给出的教程和几个示例。同时，强烈推荐学习小土堆的教程（如果想快点上手，重点推荐这个教程）。
        - PyG 是 PyTorch 专为图学习所开发的工具包，重点在于掌握基本功能，推荐学习PyTorch Geometric 教程（不断更新中） - 知乎 (zhihu.com)
    - 在有了上述深度学习原理和代码基础的前提下，对图机器学习的代码进行深入学习：图神经网络（GNN）最简单全面原理与代码实现_gnn基本原理-CSDN博客（重点关注节点分类的代码实现，教程中的模型示例为 GCN，对应论文：ICLR'17 GCN，经由 PyG 封装前的原始代码：PyTorch-GCN），同时强烈建议完成斯坦福CS224W课程对应的 Colab，参考答案：CS224W-Colab。
    - 对LaTeX不熟悉的同学请学习LaTeX基础教程。

2. 科研基础（1天左右）：通过网络搜索
    - 了解什么是 arXiv、什么是 Overleaf；
    - 了解如何进行文献管理和阅读、如何做文献笔记和论文调研报告；
    - 了解 PapersWithCode, Google Scholar, Zotero；
    - 了解 BibTeX，知道参考文献的几种写法和格式；
    - 了解计算机领域发表学术论文的会议和期刊评级机制 CCF-A/B/C（区别于传统的 SCI 分区）CCF推荐国际学术刊物目录-中国计算机学会，重点关注以下分类：网络与信息安全、数据库/数据挖掘/内容检索、人工智能、交叉/综合/新兴。

3. 文献阅读（两周左右）：这部分主要是为你打下 AI 的研究基础。需要阅读一些相关领域的经典论文，复现其中的一些未来常用的方法论文，并**完成一份学习报告**，为自己未来在研究生期间的工作打下坚实的基础。
    - 阅读报告的内容包括但不限于：
        - 第一章：概念学习。内容应当简明扼要，主要谈理解（所选择方向的研究动机，基本流程，重要概念），拒绝从网上直接复制粘贴；这部分内容不超过5页，以组内方向之一联邦图学习为例；

联邦图学习：了解联邦图学习的基础概念，为什么有联邦图学习这个需求。辨析联邦图学习的类别、横向和纵向联邦图学习。了解不同联邦图学习基础的训练流程，思考以下问题：客户端本地如何更新？是否有中心服务器？服务器的作用是什么？客户端上传到服务器或者互相通信的内容（模型、特征等..）？

        - 第二章：文献阅读报告 (注意，为了避免形式主义的任务式阅读，有个小提醒：该部分的本质是为了提高你深入阅读论文的能力，提高你对方法抽象概念的理解，一些技术细节可以不用钻牛角尖，重点还是论文背景和问题定义、论文动机和贡献解读)。

          建议根据不同的研究方向阅读以下给出的全部论文（见本入门培训最后），并在所给文章中按个人喜好**挑选 3 篇**撰写阅读报告，建议选择较新的文献。

            - 阅读报告的内容应包括：论文标题、作者单位、论文背景和问题、论文动机和贡献解读、方案设计详细分析、实验效果及其分析、结论、自己的思考等。
            - 撰写阅读报告时可以参考原论文、网络解读（博客、知乎专栏等）、跟进本论文的工作等。（注意：某些较为老旧的论文原文可能在 Introduction 和 Method 部分的写作较为晦涩，经常会出现难以理解的公式、符号等，此时不用灰心，完全可以依靠网络博客、视频等的解读来学习这部分的内容）这部分内容页数不限；
            - 注意：重点关注论文的背景、问题、动机、方法部分，实验部分适当取舍阅读，思考部分随意写一点自己真实的想法就可以（没必要上GPT来完成）

        - 第三章：代码复现报告。你需要在选择的方向的中挑选**任意 2 篇文献**（参考👆🏻 给出的文献），根据他们给出的 GitHub 仓库跑通代码，提供跑通训练 pipeline 到 test 的完整流程，并对比其是否和原文中的结果吻合。建议选择最新的文献完成代码复现。

- 第四章：文献 PPT 报告。你需要任意挑选一个方向，挑选**其中 2 篇论文（独立于文献阅读报告中的 3 篇论文）**做一个PPT对这两篇论文进行展示。阅读和展示考核文献时，你尤其需要思考：本文的创新点在哪里？未来是否还有能继续研究的空间？你的思考是什么？

- 第五章：未来展望（随意写一点真实感悟就好）。你需要在本文的全部内容的基础上，展现你在考核期间学习到的内容、所学内容的自我思考和未来展望，阐述自己在考核期间的心得和收获，并对考核任务的难度、平滑度进行点评和提出建议，同时阐明自己在研究生期间的学习目标和感兴趣的研究方向。

重要提醒：在你看论文的时候，可能会看到一些复杂公式不好理解，容易望而生畏，这里有一些经验建议：

1. 跳过公式，理解算法核心，做概念抽象的理解
2. 结合知乎或者优秀帖子看，先用中文和别人嚼碎的，会好接受一点
3. 结合代码跑着（断点调试），理解抓住数据 input 和 output，数据在整个 model 是如何流动的（forward）（关注matrix shape）

4. 文献调研报告的格式提示（尽量完成就好，也是规范日后学术PPT的标准）：
   - 推荐一级标题四号字体，二级标题和正文部分都用小四；
   - 中文一律使用微软雅黑，英文部分使用Times New Roman字体；
   - 图片、表格需要有标题。涉及到参考文献的需要设置交叉引用；
   - 可以使用Word，也可以使用LaTeX，以文档美观、易于阅读为最终目标。提交的报告文件必须是PDF格式；
   - 报告应图文并茂、排版美观；代码部分建议截图，无需粘贴代码，以美观为重。

5. PPT格式提示：
   - PPT可以用中文或英文制作；
   - 不允许设置动画，导出为PDF格式；
   - 中文一律使用微软雅黑，英文部分使用Times New Roman字体；
   - 图片、表格需要有标题。涉及到参考文献的需要设置交叉引用；参考文献放在本页PPT的最下方；
   - 可以使用Powerpoint，也可以使用LaTeX，以文档排版美观、易于阅读为最终目标。提交的报告文件必须是PDF格式。

友情提示：这个不是任务，与其说是考核，不如说是我们更想让你更加深入理解科研的组成部分，明白如何进入到科研的模式中。我们不希望出现为了应付大作业式 GPT 的思考，你的内容可以具体或者天马行空，但不要泛泛而谈。字数、内容都不作为好坏的标准，我们希望能够这个过程让你真正提升的，不仅知识水平，还有自学能力。

- 联邦学习：了解联邦学习背景、概念；了解并辨析横向联邦学习和纵向联邦学习的区别；了解 Global 和 Personalized 联邦学习的区别和联系；了解最基本的联邦学习 baseline：FedAvg，并且跑通一个demo，了解一些常用 Framework (FATE、FedML)（了解一下就可以，不用跑）。了解联邦学习中的通信、隐私保护、数据结构等（不是重点）。了解联邦学习中一些常用算法：FedProx、SCAFFOLD、FedBN，了解一些常用改进模型性能的手段：MOON、FedProto、kNN-Per、FCCL。了解组里的通用联邦学习框架。附：联邦学习资源List。

- 如果对其他方向感兴趣，比如：大语言模型时代下的图学习，强烈推荐阅读高质量的综述论文：IJCAI'24 From CUHK、KDD'24 From HKU，同时可以自行发挥，查询相关资料（可以来源于论文、Talk、知乎、微信公众号、bilibili 等常见媒介），按对等要求完成。

图学习基础（无论选择哪一研究方向，以下论文必读）

1. ICLR'17 GCN
2. NeurIPS'17 GraphSAGE
3. ICLR'18 GAT
4. ICML'19 SGC
5. ICML'20 GCNII

方向一：以数据为中心（Data-centric）的新型图神经网络架构和学习范式

高阶图数据：有向图、符号图、超图等...（二部图、时序图）

数据挑战：

- 数据质量：不平衡、噪音、分布外；
- 数据数量：标注、增强；
- 数据效率：蒸馏、压缩、选择；
- 数据隐私：遗忘、差分隐私

学习范式：

- 面向大规模数据的可扩展图学习；
- 提升数据效率的主动学习；
- 数据动态更新的持续学习；
- 数据噪音或稀疏环境下的鲁棒图学习等...

无向图：

1. ICLR'17 GCN
2. ICLR'18 GAT

有向图：

1. NeurIPS'21 MagNet
2. LoG‘24 Dir-GNN
3. VLDB'24 LightDiC
4. ICDE'24 ADPA

符号图：

5. AAAI'20 SNEA
6. KDD'21 GS-GNN
7. WWW'23 RSGNN

超图：

8. AAAI'19 HGNN
9. ICLR'22 AllSet
10. KDD’24 DPHGNN
11. ICLR’24 LightHGNN

方向二：分布式联邦图学习

1. AISTATS’17 FedAvg
2. NeurIPS'21 FedSage
3. AAAI’22 FedProto
4. ICML'23 FedPub
5. VLDB'23 FedGTA
6. ICDE'24 AdaFGL
7. IJCAI'24 FedTAD

方向三：注重隐私保护的图遗忘学习

1. CCS'22 GraphEraser
2. AISTATS'23 Projector
3. USENIX'23 GUIDE
4. ICLR'23 GNNDelete
5. ICLR'23 CGU
6. WWW'23 GIF
7. AAAI'24 MEGU

方向四：大语言模型时代下的图学习

1. ICLR'23 GLEM
2. SIGIR'23 G2P2
3. SIGIR'24 GraphGPT
4. ICLR'24 OFA
5. ICLR'24 TAPE
6. IJCAI'24 ENGINE
7. WWW'24 GraphAdapter
8. WWW'24 GraphTranslator