◁ BIT ▷

以数据为中心的图智能
Data-centric Graph Intelligence

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

分享人：李洵楷

德以明理　学以精工

概述

“以数据为中心的图智能”是指在大模型时代下，通过结构化数据建模世界实体间的复杂关系，并实现通用表征与知识挖掘的核心技术。该方向包含以数据为中心的人工智能 (Data-centric AI)、图机器学习 (Graph ML)，以及 AI4Science 等研究内容，是迈向通用人工智能 (AGI) 的关键。当下我们的研究重点与大模型紧密结合。

研究方向(数据维度)和研究视角

- 多模态属性图：通过建模具有多种模态信息的实体（节点）间关系（边），我们旨在提出多模态图基础模型并建立新型检索范式以促进实体、模态的多层次语义交互与融合，赋能于有关多模态理解与生成的实际应用
- 生命科学与智能体协作图：通过图来建模生命实体（蛋白质的空间构象、药物作用关系、细胞组学依赖）间的复杂联系和智能体协作关系，我们旨在依赖基于图和大模型的人工智能方法促进科学发现和多智能体应用
- 数据管理视角：通过将大模型视作 Generator、Selector、Embedder、Evaluator，搭建人在环外工作流，实现数据自动化生成与收集、清洗与选择、处理与增强、评估与治理，从数据管理的视角改善传统 AI 应用

近期组内研究成果

- 数据数量：
  - 数据标注 [arXiv 2025 OGA]
  - 数据合成 [arXiv 2025 GraphMaster]
  - 数据扩展（数据规模） [WWW 2024 ATP] [VLDB 2024 LightDiC] [arXiv 2025 ScaDyG]
- 数据质量：
  - 数据范式（数据类型） [TKDE 2023 AHGAE] [ICDE 2024 ADPA] [WWW 2025 MAP]
  - 数据增强 [arXiv 2025 UltraTAG] [arXiv 2025 LLaTA] [arXiv 2025 LRW-OOD]
- 数据效率：
  - 数据蒸馏（数据选择）[VLDB 2025 GEC] [ICML 2025 EDEN]
  - 数据泛化（基础模型） [arXiv 2025 MoT]

近期组内研究成果

- 数据隐私：
  - 数据联邦 [VLDB 2023 FedGTA] [IJCAI 2024 FedTAD] [ICDE 2024 AdaFGL] [IJCAI 2025 FedGM] [VLDB 2025 OpenFGL] [arXiv 2025 FedPG] [arXiv 2025 FairFGL] [arXiv 2025 FedGKC] [arXiv 2025 FedC4] [arXiv 2025 FedGFM]
  - 数据遗忘 [AAAI 2024 MEGU] [arXiv 2025 SGU] [arXiv 2025 OpenGU].
- 研究小组：
  - 图与大模型赋能多模态理解与生成

Graph-enhanced Multimodal Large Model, Multimodal Graph Retrieval Augmented Generation

  - 图与大模型助力科学发现与多智能体协作

Protein Design, Drug-Drug-Interaction, Cell Representation, Agent System

前言

近年来，在李老师的带领下，课题组专注于图数据科学领域，围绕图算法、谱图理论和图学习开兼顾理论深度和应用广度的深入研究。在师生共同努力下，课题组积累了深厚的理论基础，并取得了丰硕的研究成果。

本教程以图学习为例，覆盖从 AI 基础到科研入门的全过程，带领大家从零开始独立完成科研任务，体验完整的科研流程。目前，我们聚焦图与大模型领域，研究主题广泛，涵盖丰富的应用场景，总能找到你感兴趣的方向。针对大模型研究所需的算力需求，实验室提供了充足的计算资源。此外，学长学姐将随时为你提供答疑与指导。

加入实验室不受年级或专业限制，欢迎所有对科研充满兴趣的同学加入！

前言

这个 Learning Research Tutorial (如何从零开始做科研 – PPT Version) 会将科研经验进行书面总结，而不是以口口相传的方式进行传承，从而更好的帮助大家学习如何进行科研。

无论是新加入或志愿加入课题组的本科生还是已经加入课题组一段时间的硕士/博士研究生，相信都可以从中学到一些关于如何做科研，做好科研的方法论。

- 对于本科生，这个 Tutorial 可以帮助你 从零开始建立科研基础和科研思维。
- 对于硕士/博士研究生，这个 Tutorial 可以引导你 开展体系化的研究，做出有影响力的科研成果。

前言

更加具体来说，对于新加入或志愿加入课题组的本科生

通过这个 Learning Research Tutorial，你将收获以下内容

- 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；
- 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；
- 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；
- 转换科研项目成果，斩获各种竞赛奖项 (国家级竞赛例如挑战杯等)，大厂名组访问实习。

前言

对于加入课题组一段时间的硕士/博士研究生，相比于具体的学习路线，通过这个 Tutorial 我更希望结合自己实际经历分享一些关于做好科研的经历经验和方法论，让大家的科研之路走得更加顺畅。

个人认为，Top Ph.D. Student 懂得设定一个长远的科研目标。这个科研目标具有重要的科学价值和实际价值(在实际应用中寻找真正有价值的科学问题)。然后根据这个科研目标细化科研的 Roadmap。博士期间做的几篇论文都是围绕着解决这个科研目标，并且做的论文能够清晰地展示出自己沿 Roadmap 的科研进展。

更加具体的说，Ph.D. Student 需要有五方面的能力

(1) 寻找重要的科研问题；(2) 提出解决方案；(3) 做实验；(4) 写论文；(5) 做 Presentation。

如何入门科研

希望能在入门教程的指引下，带领你掌握以下内容

- 人工智能/机器学习/深度学习的基本原理
  - ✓ 数据集、模型、训练 (损失函数、梯度) 、推理等机器学习的基本原理。
- 深度学习代码的基本功
  - ✓ 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模型训练，能在给出 GitHub 仓库的情况下，根据论文内容跑通代码，对已有代码进行改进。
- 科研基础
  - ✓ 了解什么是 arXiv、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。

如何入门科研

深度学习的基本原理和代码功底（两周左右）：

基本原理

a. 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，其余内容暂时不用学习。

b. 参考 斯坦福 CS224W 课程，学习图机器学习的基本原理：

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，其余内容暂时不用学习。

如何入门科研

深度学习的基本原理和代码功底（两周左右）：

代码能力：

a. 搭建本地代码环境，强烈推荐使用 VSCode + Anaconda 进行管理，关于如何在本机配置开发环境，可参考 配置教程；

b. 如果对 Python 语言没有太多了解，建议学习 (1) Python 基础教程；(2) CS231n 教程；(3) Jupyter Notebook 教程；

c. 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 PPT Slides) 课程的基础上，学习以下给出的教程和几个示例 PyTorch 官方文档；

d. PyTorch Geometric (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 PyG 教程；

e. 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，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。

如何入门科研

科研基础（3天左右）：

a. 对 LaTeX 不熟悉的同学请学习 LaTeX 基础教程 和 8分钟入门 Overleaf + LaTeX；

b. 对 Linux 不熟练的，学习 Linux 基础教程；

c. 对 Git 不熟悉的，学习 一小时 Git 教程；

d. 通过网络搜索，了解什么是 arXiv、什么是 Overleaf；

e. 通过网络搜索，了解 PapersWithCode, Google Scholar, Zotero，初步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；

f. 可以考虑学习使用 PowerPoint，或者 LaTeX 的 Beamer 包完成汇报 PPT 的制作；

g. 了解 BibTeX，知道参考文献的几种写法和格式。

科研进阶

论文阅读（2周左右）：

在掌握机器学习、深度学习的基本原理后，选择一个🌟代表的具体方向，阅读该方向所罗列的基础论文

该阶段可以略微放宽对于论文理解的要求，“不求甚解”，切忌对复杂的理论推导、模糊的实现细节钻牛角尖取而代之的，我们希望你可以做到：

- 对论文中提到的陌生术语或技术，进行针对性检索，大致掌握其含义，不必拘泥于背后的数学理论或实现细节
- 系统性的建立自己对该领域的理解和认识，即能够提取某一具体方向内所有基础论文的共性，以自己的方式对他们的研究内容进行总结
- 关注论文的基础实验设置，掌握数据集信息、基线方法和性能评估策略等

科研进阶

研究方向一：该方向围绕多模态属性图 (Multimodal Attributed Graph, MAG) 展开，旨在依托预训练大模型 (Pre-trained Large Model, PLM) 实现 PLM4Graph 和 Graph4PLM

从 PLM4Graph 的角度来说，我们希望依赖 PLM 对 MAG 的多模态语义进行统一投影，同时借助其生成能力，构造具备通用性的下一代图基础模型，即图增强的多模态大模型，能够同时支持图相关和多模态相关的下游任务。

从 Graph4PLM 的角度来说，我们希望构造以多模态属性图为核心的知识数据库，基于提问实时的从中进行高效的信息检索，将检索内容与提问进行拼接作为 prompt 送入下游任意 PLM 中以避免幻觉问题。

核心技术词条：Graph Neural Network，Multimodal Large Model，Multimodal Representation Learning，Self-supervised Representation Learning，Retrieval Augmented Generation，Knowledge Graph，Information Retrieval

科研进阶

研究方向一：该方向围绕多模态属性图 (Multimodal Attributed Graph, MAG) 展开，旨在依托预训练大模型 (Pre-trained Large Model, PLM) 实现 PLM4Graph 和 Graph4PLM

PLM4Graph, Next-generation Graph Foundation Model 🌟

i.e., Graph-enhanced Multimodal Large Model

1. [KDD'25 MAGB] https://arxiv.org/abs/2410.09132
2. [WWW'25 UniGraph2] https://arxiv.org/abs/2502.00806
3. [CVPR'25 GRAPHGPT-O] https://arxiv.org/abs/2502.11925
4. [CVPR'25 MM-Graph] https://arxiv.org/abs/2406.16321
5. [arXiv'25 NTSFormer] https://arxiv.org/abs/2507.04870

Graph4PLM 🌟

i.e., Multimodal-enhanced GraphRAG

1. [EMNLP'22 MuRAG] https://arxiv.org/abs/2210.02928
2. [SIGIR'25 MRAMG-Bench] https://arxiv.org/abs/2502.04176
3. [arXiv'25 PathRAG] https://arxiv.org/abs/2502.14902
4. [ACL'25 HopRAG] https://arxiv.org/abs/2502.12442
5. [VLDB'25 Survey] https://arxiv.org/abs/2503.04338

科研进阶

研究方向二：该方向围绕生命科学与智能体协作图展开，虽然是两个截然不同的方向，但是它们的核心动机都是通过图建模技术来挖掘实体间的复杂关联，从而实现 Graph&PLM4Application （生命科学与多智能体应用）

对于生命科学来说，我们主要关注 AI4Protein，AI4Drug，AI4Cell，通过引入图技术，明确建模多样化生命实体间的复杂关联，从而实现知识挖掘，以计算密集型的依赖大模型的人工智能方法，更好地完成相关的下游任务

对于多智能体系统来说，我们将 PLM 视作智能体节点，通过引入图来明确建模多智能体间的协同调度关系，针对某一具体的应用场景，依赖图学习方法获取高效调度决策和智能体剪枝等优化方案

核心技术词条：Graph Neural Network，Pre-trained Large Model， Multimodal Representation Learning， Science Foundation Model，Self-supervised Representation Learning，Graph Structure Learning，Auto Workflow

科研进阶

研究方向二：该方向围绕生命科学与智能体协作图展开，虽然是两个截然不同的方向，但是它们的核心动机都是通过图建模技术来挖掘实体间的复杂关联，从而实现 Graph&PLM4Application （生命科学与多智能体应用）

Graph&PLM4Science

e.g.., AI4Protein

1. [NeurIPS'24 S3F] https://arxiv.org/abs/2412.01108
2. [ICML'24 Surface-VQMAE] https://proceedings.mlr.press/v235/wu24o.html

e.g.., AI4Drug

3. [AAAI'24 TIGER] https://ojs.aaai.org/index.php/AAAI/article/view/27777
4. [AAAI'24 MKG-FENN] https://ojs.aaai.org/index.php/AAAI/article/view/28887

e.g.., AI4Cell

5. [NeurIPS'24 ScCello] https://arxiv.org/abs/2408.12373
6. [NM'24 ScFoundation] https://www.nature.com/articles/s41592-024-02305-7

Graph&PLM4Agent

1. [ICML'25 G-Designer]
https://arxiv.org/abs/2410.11782
2. [NeurIPS'24 GNN4TaskPlan]
https://arxiv.org/abs/2405.19119
3. [arXiv'25 Survey]
https://arxiv.org/abs/2506.18019

分享完毕
欢迎提问交流

Thanks for your listening

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY