◁ BIT ▷

# 从零开始做科研：基础到进阶的全面教程
# ---以大模型与图学习为例

分享人：李洵楷

时间：2024年11月28日

# 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。

本教程以图机器学习为例，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，全面体验完整的科研流程。

目前，Graph Learning 小组聚焦图学习与大模型领域，研究主题广泛，涵盖丰富的应用场景，总能找到契合你兴趣的方向。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供大家使用。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硕士研究生、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。

- 对于硕士研究生、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硕士研究生、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。

# 1. 如何入门 LLM&Graph 的科研



入门教程的学习时长大约为一个月，希望能在培训任务的指引下，带领你掌握以下内容

- 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。
- 图机器学习
  - ✓ 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)。
- 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。
- 基本的 PPT 制作技能
  - ✓ 会使用 PowerPoint，或者 LaTeX 的 Beamer 包完成 PPT 的制作;

# 1. 如何入门 LLM&Graph 的科研



1. 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：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，其余内容暂时不用学习。

# 1. 如何入门 LLM&Graph 的科研



## 1. 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。

# 1. 如何入门 LLM&Graph 的科研

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

## 2. 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。

# 1. 如何入门 LLM&Graph 的科研



## 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。需要阅读未来研究领域的经典论文，并完成一份文献阅读报告，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，拒绝从网上直接复制粘贴；
  - ✓ 以组内方向之一联邦图学习为例：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。了解联邦图学习基础的训练流程。
- 第二章：文献阅读报告
  - ✓ 你需要在选择的方向的中挑选任意三个文献撰写阅读报告，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。

# 1. 如何入门 LLM&Graph 的科研



- 第三章：代码复现报告
  - ✓ 你需要在选择的方向的中挑选任意两个文献，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：考核文献阅读报告
  - ✓ 你需要在选择的方向的中挑选任意两个文献，完成文献阅读报告，并做一个 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。

# 1. 如何入门 LLM&Graph 的科研



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 (断点调试)，抓住 Input 和 Output，数据在 Model 中如何流动 (Forward) (关注 Matrix Shape)。

- 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；
- 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，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、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 P14-P16)。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)。

# 1. 如何入门 LLM&Graph 的科研



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](https://arxiv.org/abs/1609.02907)
2. [NeurIPS'17 GraphSAGE](https://arxiv.org/abs/1706.02216)
3. [ICLR'18 GAT](https://arxiv.org/abs/1710.10903)
4. [ICML'19 SGC](https://arxiv.org/abs/1902.07153)
5. [SIGIR'23 G2P2](https://arxiv.org/abs/2305.03324)
6. [ICLR'24 OFA](https://arxiv.org/abs/2310.00149)
7. [ACL'24 InstructGraph](https://arxiv.org/abs/2402.08785)
8. [NeurIPS'24 GFT](https://arxiv.org/abs/2411.06070)
9. [NeurIPS'24 ProG](https://arxiv.org/abs/2406.05346)
10. [arXiv'24 GFSE](https://openreview.net/pdf?id=JQT6iGrXTh)
11. [WWW'25 GraphCLIP](https://arxiv.org/abs/2410.10329)
12. [WWW'25 SAMGPT](https://arxiv.org/abs/2502.05424)

# 1. 如何入门 LLM&Graph 的科研



## 方向一 开放世界环境下以数据为中心（Data-centric）的通用图学习

近年来，人工智能的发展面临挑战，因为许多领先的 LLM 仍然依赖于 Transformer 架构，性能提升已从模型转向以数据为中心的策略。

对于图学习来说，结构化数据的元数据涌现和学习范式也逐渐在大模型的参与下向以数据为中心的角度倾斜。课题主要考虑在 LLM 的协助下，解决开放世界环境下的一系列数据挑战，迈向更加通用、实用、可部署的图智能

(1) 开放世界环境下的数据挑战：动态流图数据下所产生的新旧数据差异：特征偏移、标签偏移、拓扑偏移、类型偏移；

(2) 以数据为中心的图学习范式：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、数据噪音或稀疏环境下的鲁棒图学习等；

1. [AAAI'21 ER-GNN](https://arxiv.org/abs/2003.09908)
2. [AAAI'21 TWP](https://arxiv.org/abs/2012.06002)
3. [arXiv'24 LEGO-Learn](https://arxiv.org/abs/2410.16386)
4. [CoLLAs'24 POWN](https://arxiv.org/abs/2406.09926)
5. [ICDE'24 OpenIMA](https://arxiv.org/abs/2403.11483)
6. [NeurIPS'24 KG-FIT](https://arxiv.org/abs/2405.16412)
7. [NeurIPS'24 TPP](https://arxiv.org/abs/2410.10341)
8. [MM'24 FTF-ER](https://arxiv.org/abs/2407.19429)

# 1. 如何入门 LLM&Graph 的科研



## 方向二 基于大语言模型的可信图（Trustworthy）学习

“万物为图”即任意的关系型数据都可以建模为图，从社交网络到推荐系统，图在各个领域无处不在。尽管引入大语言模型的图神经网络（图基础模型）已经成为当下图分析的主流工具，但是仍存在各种可信性问题例如信息泄露、分类偏差和缺乏可解释性。本课题主要考虑定义并解决 **LLM** 赋能下的图基础模型所存在的潜在可信性问题：

(1) 联邦学习：基于分布式的协作框架，开发大语言模型所赋能联邦图学习的新技术与新方法；

(2) 遗忘学习：基于遗忘学习范式，探索解决图基础模型中可信性问题的新方案。

1. [NeurIPS'21 FedSage](https://arxiv.org/abs/2106.13430)
2. [ICML'23 Fed-Pub](https://arxiv.org/abs/2206.10206)
3. [VLDB'23 FedGTA](https://arxiv.org/abs/2401.11755)
4. [ICDE'24 AdaFGL](https://arxiv.org/abs/2401.11750)
5. [IJCAI'24 FedTAD](https://arxiv.org/abs/2404.14061)
6. [CCS'22 GraphEraser](https://arxiv.org/abs/2103.14991)
7. [ICLR'23 GNNDelete](https://arxiv.org/abs/2302.13406)
8. [WWW'23 GIF](https://arxiv.org/abs/2304.02835)
9. [AAAI'24 MEGU](https://arxiv.org/abs/2401.11760)
10. [arXiv'25 OpenGU](https://arxiv.org/abs/2501.02728)

# 1. 如何入门 LLM&Graph 的科研



## 方向三 AI4Science [图 + 大模型]

图结构为建模和表征实体间的相互作用提供了一种自然且有效的方式，而 LLM 凭借其强大的推理与泛化能力最近收到各界的广泛关注。这两者在 AI4Science 领域发挥着至关重要的作用，例如用于表征蛋白质中氨基酸链在空间中的复杂几何构象、解析细胞间的相互作用机制以及模拟生物或物理系统中的通讯网络。本课题主要考虑**蛋白质大模型**的相关研究

1. [ICLR'21 IEConv](https://arxiv.org/abs/2007.06252)
2. [PEDS'23 MIF](https://www.biorxiv.org/content/10.1101/2022.05.25.493516v1)
3. [ICLR'23 GearNet](https://arxiv.org/abs/2203.06125)
4. [NB'23 FoldSeek](https://www.nature.com/articles/s41587-023-01773-0)
5. [ICLR'24 SaPort](https://www.biorxiv.org/content/10.1101/2023.10.01.560349v2)

# 分享完毕
# 欢迎提问交流

Thanks for your listening


北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY