

林睿

邮箱: rulin0212@gmail.com • 电话: (+86) 15623049887 / (+852) 57492271 • 个人主页: rlin27.github.io

教育背景

香港大学	2018.09 – 2022.09
博士, 电子电气工程学院	
毕业论文题目: <i>Novel Compression Techniques for Compact Deep Neural Network Design</i>	
武汉大学	2014.09 – 2018.06
理学学士, 数学与统计学院 (专业: 统计学)	
GPA: 3.52/4.00	

工作经历

华为香港研究所	2022.12 – 至今
研究员, AI 框架与数据结构实验室	
○ GTN-F: A General Tensor-Native Format Representation of Multimodal Data (GTN-F: 通用张量多模态数据表达)	2023.06 – 2024.07
项目概览:	
➢ 华为 0-1 技术创新研究项目。常规大模型场景的多模态数据处理流程有数据缺乏统一表达、多系统多份、难管理溯源等问题。GTN-F 旨解决上述问题, 提高大模型场景的数据存取与管理效率。	
➢ 能将来源复杂、结构异构的多模态数据表征为通用的张量数据, 以便统一的存储、管理和分析。基于 chunk 的文件存储方式, GTN-F 支持使用 idx 快速访问关联同一样本的信息, 提升 2 倍随机访问效率。	
➢ 提供张量化查询与物化视图加载功能, 方便数据过滤; 提供类似 git 的操作命令, 方便数据版本管理。	
主要贡献:	
➢ 完成数据和张量的表达及管理, 和 chunking 策略的开发。基础功能包括 schema 设计、dataset 与 tensor 对象创建、数据添加等; 进阶功能包括 dataset 信息总结、数据筛选、数据更新, 和重新分配 chunk 大小等, 并给出详细说明文档。	
➢ 针对 chunking 的存储形式, 在图基准数据集与 baselines 数据格式相比, data-driven chunking 策略能降低 80% 的数据读取与转换的开销; workload-driven chunking 策略能辅助高效并行计算效率提升 2 倍。	
➢ 完成 GTN-F 在大模型和大数据量下应用场景的可行性调研, 给出应用方案。	
○ GTN: A General Tensor-Native Data Processing Framework (GTN: 通用张量原生数据处理框架)	2022.12 – 2023.07
项目概览:	
➢ 华为 0-1 技术创新研究项目。GTN 旨在将数据科学计算任务基于张量抽象化表达, 利用张量抽象化基于异构硬件的加速优化功能, 实现运算能力的提高, 简化数据科学任务的开发、部署和优化。	
➢ 通过 GTN, 大规模的数据特征工程任务的“算”可转换为 Tensor (多维数组) + Tensor Operators (代数计算)。经过验证, 目标任务可在异构硬件上进行部署, 达到加速优化: 与 CPU 通用 baseline 相比, 可在 70% 的验证场景中, 达到 100 倍的效率提升。	
主要贡献:	
➢ 负责传统机器学习分支。对传统机器学习模型进行张量化转换的文献综述调研, 完成 GTN 机器学习分支的开发和优化。目前, GTN-ML 可将超过 50 个常用 ML 模型 (包括树类模型, 支持向量机, 聚类模型等) 转化为张量表达的模型。	
➢ 针对机器学习分支, 给出一定的数学理论支撑, 分析模型张量化的可行性及模型复杂度。	
➢ 转换后的传统机器学习模型表现相对 Sklearn 有稳定提升, 和 HummingBird 相近。部分模型提升较大, 如支持向量机类模型在数据量较大时比 Sklearn 快 50 倍以上, 比 HummingBird 快 15 倍以上。	

部分论文

期刊.....	
○ Lin, R. *, Li, C. *, Zhou, J., Huang, B., Ran, J., Wong, N. (2023). <i>Lite it Fly: An All-Deformable-Butterfly Network</i> . Brief Paper in the IEEE Transactions on Neural Networks and Learning Systems (TNNLS).	
○ Mao, R., Wen, B., Arman, K., Zhao Y., Ann Franchesca, L., Lin, R., Wong, N., Michael, N., Hu, X., Sheng, X.,	

Catherine, G., John Paul, S. & Li, C. (2022). *Experimentally Realized Memristive Memory Augmented Neural Network*. Nature Communications.

- Tao, C.*, **Lin, R.***, Chen, Q., Zhang, Z., Luo, P., & Wong, N. (2022). *FAT: Learning Low-Bitwidth Parametric Representation via Frequency-Aware Transformation*. IEEE Transactions on Neural Networks and Learning Systems (TNNLS).

会议.....

- Ran, J., **Lin, R.**, Li, C., Zhou, J., Wong, N. (2023). *PECAN: A Product-Quantized Content Addressable Memory Network*. Design, Automation and Test in Europe Conference (DATE'23).
- **Lin, R.***, Ran, J. *, Chiu, K.H., Chesi, G., Wong, N.* (2021). *Deformable Butterfly: A Highly Structured and Sparse Linear Transform*. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS'21).
- **Lin, R.***, Ran, J.*, Wang, D., Chiu, K. H., & Wong, N. (2021). *EZCrop: Energy-Zoned Channels for Robust Output Pruning*. In proceeding of the Winter Conference on Applications of Computer Vision (WACV'22).
- Cheng, Y., **Lin, R.**, Zhen, P., Hou, T., ... & Wong, N. (2021). *FASSST: Fast Attention Based Single-Stage Segmentation Net for Real-Time Instance Segmentation*. In proceeding of the Winter Conference on Applications of Computer Vision (WACV'22).
- Ko, C. Y., **Lin, R.**, Li, S., & Wong, N. (2019). *MiSC: Mixed Strategies Crowdsourcing*. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence Main track (IJCAI'19) (pp. 1394-1400).

* 共同作者声明

曾获奖项

港研之星（团队） 华为香港研究所	2024
港研之星（个人） 华为香港研究所	2023
黑客松软件挑战赛优胜奖（前 5%） 华为	2023
数学与统计学院中法班奖学金 武汉大学	2015, 2016, 2017
数学/跨学科建模竞赛优异奖 The Consortium for Mathematics and its Application (COMAP)	2017
新生入学三等奖学金 武汉大学	2014

其他

- **编程语言:** Python/MATLAB/Git (有项目经验), SQL (熟悉)
- **语言:** 普通话 (母语), 英语 (可作为工作语言), 粤语 (熟练)
- **证书:** AWS Certified Database – Specialty (过期日期: 2027.03.23)