Lin Rui

Email: ruilin0212@gmial.com • **Homepage**: rlin27.github.io

EDUCATION

The University of Hong Kong **Sept. 2018 – Sept. 2022**

Ph.D. in the Dept. of Electrical and Electronic Engineering

Thesis Title: Novel Compression Techniques for Compact Deep Neural Network Design

Wuhan University **Sept. 2014 – Jun. 2018**

B.S. in the School of Mathematics and Statistics (Major: Statistics)

GPA: 3.52 / 4.00

WORKING EXPERIENCE

Huawei Hong Kong Research Center **Dec. 2022 – Present**

Researcher in the AI Framework & Data Tech. Lab.

- **Data Curation: Pretrain and SFT Data Preparation for Large Text-to-Video Generation Models and the Design of Tag System for Multimodal Data** **Jul. 2024 – Present**

  Project Overview:

  - This work aims to boost the performance of Xiaoyi, who is Huawei's AI agent.
  - For data preparation, the primary challenges are: (a) determining the proper ratio of data belonging to different categories, (b) maintaining the diversity of the data while making the semantics concentrated, and (c) improving the efficiency of data selection by utilizing less labor resources.
  - For tag system design, the goal is to develop a system that can cover Xiaoyi's application scenarios, while taking the correlation between different modal data into consideration.

  Main Contributions:

  - Developed a deduplication pipeline for the pretrain data and an interactive data selection tool.
  - Finished designing the tag system, which has around 40 level-1 labels and 380 level-2 labels. The test labeling accuracy of Xiaoyi's application scenarios reaches above 90% for both level-1 and level-2 labels.
- **GTN-F: A General Tensor-Native Format Representation of Multimodal Data** **Jul. 2023 – Jul. 2024**

  Project Overview:

  - 0-1 Technology & Research Innovation Project in Huawei. This project aims to solve the problems in large model scenarios like (a) the lack of unified data representation, (b) multiple copies are required due to multiple systems, and (c) difficulties in management and source tracing, thus improving data access and management efficiency.
  - GTN-F can convert multi-modal data from different sources with heterogeneous structures into the general tensor format. Based on the chunk storage strategy, GTN-F supports using idx to quickly access all necessary information related to the same sample, improving the random-access efficiency by 2×.
  - GTN-F provides tensor query and materialized view loading functions to facilitate data filtering, and operation commands similar to git are provided to facilitate data version management.

  Main Contributions:

  - Completed developing and optimizing the data format and chunking strategy and wrote detailed API documents. Basic functions include Dataset and Tensor object creation, data extension, etc. Advanced functions include dataset information summary, data filtering, update, and chunk size reallocation etc.
  - For the chunking strategy, for the graph benchmark dataset and data format: (a) the data-driven strategy can reduce 80% of the cost of loading and transforming the data, and (b) the workload-driven strategy boosts the performance of the parallel computing by 2×.
- **GTN: A General Tensor-Native Data Processing Framework** **Dec. 2022 – Jul. 2023**

  Project Overview:

  - 0-1 Technology & Research Innovation Project in Huawei. This project aims to leverage tensor abstraction as the basis for automatic optimization and enable the compilation of the same code over heterogeneous hardware.

- With GTN, the computation process of large-scale feature engineering tasks can be converted into Tensor plus Tensor Operators.
- By extensive experiments, it is verified that target tasks can be deployed on heterogeneous hardware to achieve acceleration optimization: compared with the common CPU baseline, the efficiency can be improved by 100× in 70% verification scenarios.

Main Contributions:

- Completed the development and optimization of the traditional machine learning branch in GTN. Currently, GTN-ML can convert more than 50 widely used traditional ML models (including tree-like models, SVMs, regression models, etc.) to their tensor-representation counterparts.
- The converted models perform better than the Sklearn implementation and have comparable results with the HummingBird implementation. Some models have significant performance improvements, for example, SVMs can be more than 50x faster than Sklearn implementation and more than 15× faster than HummingBird implementation when the dataset is large.

SELECTED PUBLICATIONS

JOURNAL……...

- **Lin, R.** *, Li, C. *, Zhou, J., Huang, B., Ran, J., Wong, N. (2023). *Lite it Fly: An All-Deformable-Butterfly Network*. 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).

CONFERENCE……

- 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).

* Equal Authorship Statement

HONORS & AWARDS

Star of AI Framework & Data Tech. Lab (Team) **2024**
Huawei Technologies Co., Ltd.

Star of AI Framework & Data Tech. Lab (Individual) **2023**
Huawei Technologies Co., Ltd.

Award of Excellence in the Hackathon Software Challenges Contest (top 5%) **2023**
Huawei Technologies Co., Ltd.

Sino-Francaise Scholarship in the School of Mathematics and Statistics **2015, 2016, 2017**
Wuhan University

Meritorious Winner in Mathematical/Interdisciplinary Contest in Modeling **2017**
The Consortium for Mathematics and its Application (COMAP)

Third-Class Scholarship for the First-year Freshmen **2014**

Wuhan University

ADDITIONAL

- **Programming Languages:** Python/MATLAB/Git (experienced in), SQL (familiar with)
- **Languages:** Mandarin (native), English (full working proficiency), Cantonese (good command)
- **Certification:** AWS Certified Database – Specialty (expired on 2027.03.23)