Qingquan (QQ) Song

Tel: +1 (979) 422-2777 Email: ustcsqq@gmail.com Google Scholar LinkedIn

RESEARCH INTERESTS

My research interests lie broadly in the foundation and applications in efficient ML currently focusing on:

- LLM performance optimization and compression: distillation, quantization, pruning and alignment;
- LLM for RecSys with efficient model and system co-design;
- Automated machine learning;

EDUCATION AND INDUSTRY EXPERIENCES

LinkedIn Corporation. Mountain view, CA *June 2021 – Present*

Staff Machine Learning Engineer @ Core AI Optimization Team

- Lead the Linkedin ELLA (efficient LLM algorithm) project for (1) multiple mini model creation with **distillation, alignment, pruning, quantization**. <8B model matches the internal founcation RecSys 100B+ model performance on all major vertical tasks and (2) serving optimization with vLLM / SGLang / TensorRT LLM focusing on long-context model performance optimization and KV cache size reduction, etc. (3) training optimization with Triton kernel fusion and FSDP2 FP8 training. [Pub1 [Pub2]
- Core member of the **Liger-Kernel** project, developing efficient **Triton kernels** for LLM training. The library is being adopted across industry and academia to reduce training latencies and GPU memory consumption by over 60%. [Github] [Pub]
- Core member of the **360Brew foundation RecSys model** team, working on continuous pre-training (>500GPUs) and post-training (SFT and alignment) of the 140 billion parameter LLM with upcycled MoE architecture for ranking and retrieval tasks across LinkedIn. [Pub]
- Embedding based retrieval (EBR) and GPU serving: productionized continual learning with mixed negative sampling for EBR with 1 bit post-training quantization. Implemented custom TensorFlow and PyTorch operators for KNN with attributed matching with quantized embeddings. [Pub1] [Pub2]
- Large personalized ranking model: core designer and implementer of the first LinkedIn Feed ranking and Ads CTR billion scale deep learning models with embedding table hashing, quantization and sharding with 3D parallel training in TensorFlow. [Pub]

LinkedIn Corporation. Mountain view, CA *May 2020 – Aug 2020*
Research Intern @ Linkedin AI Algorithm Foundation
Project: Tree-Based Transferable Hyperparamter Tuning [Patent]

Facebook Inc. Menlo Park, CA *May 2019 – Aug 2019*
Research Intern @ Facebook AI Applied Research
Project: Neural Architecture Search for CTR prediction [KDD' 20]

Texas A&M University College Station, TX *Aug 2016 – May 2021*
Ph.D. in Computer Science
Advisor: Prof. Xia (Ben) Hu
Automated Recommender Systems [Thesis]
Automated Machine Learning in Action (Manning Publication) [Book]
Auto-Keras: An Efficient Neural Architecture Search System [Pub] [Github]

University of Science & Technology of China (USTC) Hefei, China *Sep 2012 – Jun 2016*
B.S. in Statistics

SELECTED PUBLICATIONS

[ArXiv 25] AlphaPO–Reward shape matters for LLM alignment. *ArXiv 2025.*

[ArXiv 24] Liger Kernel: Efficient Triton Kernels for LLM Training. *ArXiv 2024.* [code]

[KDD' 24] LiRank: Industrial Large Scale Ranking Models at LinkedIn. In *Proceedings of 30th SIGKDD Conference on Knowledge Discovery and Data Mining.*

[CIKM' 24] LiNR: Model Based Neural Retrieval on GPUs at LinkedIn. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*

[ArXiv 23] QuantEase: Optimization-based Quantization for Language Models. *ArXiv 2023.*

[OPT 23] Improved Deep Neural Network Generalization Using m-Sharpness-Aware Minimization. *NeurIPS Optimization Workshop 2022.*

[Book 22] **Qingquan Song**, Haifeng Jin, and Xia Hu. Automated Machine Learning in Action. *Manning Publications 2022.*

[Frontiers' 22] Kaixiong Zhou, **Qingquan Song**, Xiao Huang, and Xia Hu. Auto-GNN: Neural Architecture Search of Graph Neural Networks.

[KDD' 22] Generalized Deep Mixed Models. In *Proceedings of the 28th SIGKDD Conference on Knowledge Discovery and Data Mining.*

[WWW' 22] Xiaotian Han, Zhimeng Jiang, Ninghao Liu, **Qingquan Song**, Jundong Li, Xia Hu. Geometric Graph Representation Learning via Maximizing Rate Reduction. *The Web Conference 2022.*

[PhD Thesis] **Qingquan Song**. Automated Recommender Systems. *Texas A&M University. 2021*

[KDD Exp' 21] Yiwei Chen, **Qingquan Song**, and Xia Hu. Techniques for Automated Machine Learning. In *ACM SIGKDD Explorations Newsletter*

[NeurIPs' 20] Zirui Liu, **Qingquan Song**, Kaixiong Zhou, Ting-Hsiang Wang, Ying Shan, and Xia Hu. Towards Interaction Detection Using Topological Analysis on Neural Networks. In *Advances in neural information processing systems.*

[KDD' 20] **Qingquan Song**, Dehua Cheng, Eric Zhou, Jiyan Yang, Yuandong Tian, and Xia Hu. Towards Automated Neural Architecture Discovery for Click-Through Rate Prediction. In *Proceedings of the 26th SIGKDD Conference on Knowledge Discovery and Data Mining.*

[TKDD' 19] **Qingquan Song**, Hancheng Ge, James Caverlee, and Xia Hu. Tensor completion algorithms in big data analytics. In *ACM Transactions on Knowledge Discovery from Data*

[KDD' 19] Haifeng Jin, **Qingquan Song**, and Xia Hu. Auto-Keras: An Efficient Neural Architecture Search System. In *Proceedings of the 25th SIGKDD Conference on Knowledge Discovery and Data Mining.* **(Oral)**

[KDD' 19] **Qingquan Song**, Shiyu Chang, and Xia Hu. Coupled Variational Recurrent Collaborative Filtering. In *Proceedings of the 25th SIGKDD Conference on Knowledge Discovery and Data Mining.* [code]

[KDD' 18] Mengnan Du, Ninghao Liu, **Qingquan Song**, and Xia Hu. Towards Explanation of DNN-based Prediction with Guided Feature Inversion. In *Proceedings of the 24th SIGKDD Conference on Knowledge Discovery and Data Mining.* **(Oral)**

[WSDM' 18] Xiao Huang, **Qingquan Song**, Jundong Li, and Xia Hu. Exploring Expert Cognition for Attributed Network Embedding. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining.*

[KDD' 17] **Qingquan Song**, Xiao Huang, Hancheng Ge, James Caverlee, and Xia Hu. Multi-Aspect Streaming Tensor Completion. In *Proceedings of the 23rd SIGKDD Conference on Knowledge Discovery and Data Mining.* **(Oral)**