Di Liu

diliu@buaa.edu.cn · (+86) 188-012-78568 · Beijing, China

Education

Beihang University (BUAA) **M.Eng.** 09/2021 – Expected 01/2024

- Research interests: cloud/mobile computing, machine learning systems, edge intelligence, etc.

Beijing Jiaotong University (BJTU) **B.Eng.** 2017.09 – 2021.06

- GPA: 3.77/4.0, Ranking: 19/273.

Publications

- **Liu D**, Ma Z, Zhang A, et al. MagicBatch: An Energy-Aware Scheduling Framework for DNN Inference on Heterogeneous Edge Servers in Space-Air-Ground Computation[C]//DataCom, 2022.
- **Liu D**, Ma Z, Zhang A, et al. Efficient GPU Resource Management under Latency and Power Constraints for Deep Learning Inference[C]//IEEE MASS, 2023.
- Ma Z, Li Y, **Liu D**, et al. CHESS: Joint Energy and Makespan Optimization for Dynamic CNN Task Scheduling on the Heterogeneous System [C]//ICMLC, 2023.
- **Liu D**, Zhang A, Zheng K, et al. Energy-efficient Computation Offloading and Resource Allocation for Deep Learning Inference in Mobile Edge Computing [C]//IEEE INFOCOM, under review, 2024.

Reserach Experience

Energy-efficient DNN Inference Workloads Offloading in Edge Intelligence. 12/2022 – 04/2023

For the edge intelligence scenario, we model the latency and energy consumption when the system cooperatively executes DNN inference tasks and design an energy-efficient scheduling algorithm to minimize the total energy consumption of the system under the long-term average latency constraint.

Efficient GPU Server Resource Management for DNN Inference Workloads. 05/2022 – 11/2022

For the scenario where GPU servers provide DNN inference services, we measure the impact of batch size, GPU frequency and GPU spatial sharing on the performance of inference workloads, then we design an efficient resource management framework to maximize the total throughput under the constraints of latency and power.

Energy-efficient DNN Inference Scheduling on Heterogeneous Edge Servers. 12/2021 – 04/2022

For the scenario where servers are deployed with heterogeneous computing resources, we measure the latency and energy consumption characteristics of various DNN models. And we design an energy-aware scheduling algorithm to minimize the total energy consumption of the system while ensuring the upper limit of latency.

Deep Representation of Trajectory Based on Spatial-Temporal Fusion. 02/2021 – 05/2021

We encode and fuse the spatial-temporal information of the vehicle's trajectory and use the RNN model to learn from them to obtain a deep representation of the trajectory, and we also design various methods to evaluate the effectiveness of representation results.

Work Experience

Tencent Company **Software Development Engineer Intern** 06/2023 – Present

I complete the query analysis service for search advertising scenarios in Tencent Ads, including link keyword segmentation, semantic correction, and semantic analysis to improve advertising effectiveness. And I improve the performance of service by asynchronous parallel computing and distributed computing.

Others

- Honors: Academic Scholarship of BJTU and BUAA, Social Work Scholarship of BJTU, etc.
- Aawrds: Collegiate Innovation and Entrepreneurship Project, the National Award, etc.
- Skills: C++, C, Python MATLAB, MySQL, Docker, PyTorch, TensorFlow, Git, LaTeX, Origin, etc.