Xinyue Ma

PhD Student

xinyuema@unist.ac.kr
+821042470039

RESEARCH INTERESTS

- Efficient Machine Learning Systems
- Continual Learning
- On-device Learning

EDUCATION

Ulsan National Institute of Science and Technology — Feb. 2022 - Present
Combined Master and PhD in Computer Science Engineering
Affiliation: *OMNIA Lab of UNIST*
Advisor: Myeongjae Jeon

Korea Advanced Institute of Science and Technology, — Feb. 2017 - Aug. 2021
B.S. in Electrical Engineering | Minor: Computer Science
Advisor: Euijong Steven Whang

PUBLICATIONS

1. Cost-effective On-device Continual Learning over Memory Hierarchy with Miro — Oct. 2023
Xinyue Ma, Suyeon Jeong, Minjia Zhang, Di Wang, Jonghyun Choi, Myeongjae Jeon — **ACM MobiCom**

2. **(under review)** REP: Resource-Efficient Prompting for On-device Continual Learning — 2024
Sungho Jeon, **Xinyue Ma**, Kwang In Kim, Myeongjae Jeon — **CVPR**

RESEARCH PROJECTS

Automatic 3D-parallelization for Deep Learning with Heterogeneous GPUs
Collaboration with **Samsung Research*

Jun. 2023 - Present

With the rapid advancement of increasingly powerful GPUs each year, the necessity of training large-scale models on clusters with heterogeneous GPUs becomes inevitable. However, existing parallelization schemes fail to fully capitalize on the potential of this heterogeneity that stems from diverse combinations of GPUs available in the cluster, thus missing an opportunity for achieving more efficient parallelization. We aim to develop a framework that optimizes LLM training through the cost-effective exploration of potential 3D-parallelization strategies, which combine pipeline, tensor, and data parallelism on the available GPU servers. This framework will efficiently identify the optimal parallelization strategy in a timely manner by promptly solving a sophisticated cost function.

RESEARCH PROJECTS

Cost-effective On-device Continual Learning Systems

** Collaboration with **DeepSpeed** and **Yonsei University***

Apr. 2023 - Present

DL applications on the edge often require training DNN models 1) within the edge device to preserve privacy and 2) continually as data arrives, an emerging learning paradigm for edge devices known as Continual Learning. With limited resources on edge devices, cost-efficiency becomes critical for a training job. However, prior studies are predominantly simulation-based and do not have adequate consideration of cost-efficiency. Together with my colleagues, I focus on system solutions for cost-effective on-device CL, achieving high model accuracy without compromising energy efficiency, to better suit the energy-sensitive edge devices. My recent work, Miro, achieves this goal by dynamically optimizing various CL configurations such as memory and I/O usage based on resource states and exploiting the most favorable accuracy-energy trade-offs for high cost-effectiveness. Recent advances in CL incorporate the deployment of vision-transformers. However, the adoption of such models on the edge has been limited by the expensive computational cost and the memory footprint unable to fit in edge devices. My current work, REP (under review for CVPR 2024) optimizes the computational and memory efficiencies of vision-transformer to harvest the benefit on edge devices while maintaining theit superior accuracy. I am currently working on extending Miro to more general, realistic, and challenging scenarios, such as video analytics for AV applications.

EXPERIENCES

Internship	Intern @ Intelligent Edge and Cloud Group (2023. Nov. - 2024. Feb.)	*Microsoft Research Asia, China*
Teaching Assistant	Computer Architecture (2023 Fall) Parallel Computing (2022 Fall) Advanced Programming (2022 Spring)	*UNIST, Korea*

SKILLS

Programming Language	C/C++, Python, SQL, VDHL, Assembly
Tools	PyTorch, TensorFlow, MapReduce
Skillset	Socket Programming, Pintos, Verilog
Communication	Chinese (native), English (fluent, TOEFL 115/120), Korean (fluent, TOPIK 6/6)