

Sanha Maeng

RESEARCH ENGINEER AT MAUMAI

✉ sanha.maeng@gmail.com | 🏠 9rum.github.io | 🐙 [GitHub](#) | [LinkedIn](#) | 🎓 [Google Scholar](#)

Education

Sogang University

M.S. IN COMPUTER SCIENCE AND ENGINEERING

- Advisor: Prof. Sungyong Park

Seoul, Republic of Korea

Mar. 2021 - Feb. 2023

Kookmin University

B.S. IN COMPUTER SCIENCE AND ENGINEERING

Seoul, Republic of Korea

Mar. 2017 - Feb. 2021

Experience

MaumAI

RESEARCH ENGINEER (TECHNICAL RESEARCH PERSONNEL)

Seongnam, Republic of Korea

Mar. 2023 -

- Achieved sub-second time-to-first-token latency in Llama 3.2 1B serving on Qualcomm QCM6490 (12TOPS).
- Improved LLM serving throughput by up to **352×** using continuous batching, PagedAttention and FlashAttention.
- Accelerated talking face generation models inference by **4×** using ONNX and TensorRT, achieving up to **440×** speedup in time-to-first-frame latency with iteration-level scheduling. Further improved by **3×** with computation-I/O overlapping.

CONCAT

SOFTWARE ENGINEER INTERN

Seoul, Republic of Korea

Jun. 2020 - Sep. 2020

FlyHigh

SOFTWARE ENGINEER INTERN

Seongnam, Republic of Korea

Jun. 2019 - Aug. 2019

Publications

1. **Sanha Maeng**, Gordon Euhyun Moon and Sungyong Park, CHRONICA: A Data-Imbalance-Aware Scheduler for Distributed Deep Learning, 23rd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid '23) [paper](#) ↗

Open Source Contributions

1. Contributed to ExecuTorch regarding Qualcomm QCM6490 support, becoming one of the ExecuTorch v1.1.0 contributors [pytorch#16331](#) ↗ [release note](#) ↗
2. Contributed to Amazon Deep Graph Library (DGL) regarding distributed job launcher, becoming one of the DGL v2.0.0 contributors [dmlc#6304](#) ↗ [release note](#) ↗

Honors and Awards

- 2022 **Best Paper Award**, Korea Software Congress
- 2020 **2nd Place**, KMUCS Capstone Design Conference

KIISE

KMU