Sanha Maeng

Research engineer at MaumAI

sanha.maeng@gmail.com | 9rum.github.io | GitHub | LinkedIn | Google Scholar

Education

Sogang University — *Seoul, Republic of Korea*
M.S. in Computer Science and Engineering — *Mar. 2021 - Feb. 2023*
- Advisor: Prof. Sungyong Park

Kookmin University — *Seoul, Republic of Korea*
B.S. in Computer Science and Engineering — *Mar. 2017 - Feb. 2021*

Experience

MaumAI Inc. — *Seongnam, Republic of Korea*
Research engineer (technical research personnel) — *Mar. 2023 -*
- Accelerated talking face generation models inference by $4\times$ using ONNX and TensorRT, achieving up to $440\times$ speedup in time-to-first-frame latency with iteration-level scheduling.
- Implemented an LLM inference engine performance evaluation tool based on a Poisson process.
- Improved LLM serving throughput by up to $352\times$ using continuous batching, PagedAttention and FlashAttention.
- Built and deployed H100 cluster monitoring system using WandB.

CONCAT Inc. — *Seoul, Republic of Korea*
Software engineer intern — *Jun. 2020 - Sep. 2020*

FlyHigh Co., Ltd. — *Seongnam, Republic of Korea*
Software engineer intern — *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 Amazon Deep Graph Library (DGL) regarding distributed job launcher [dmlc#6304], became one of the DGL v2.0.0 contributors [release note].

Honors and Awards

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