

Sanha Maeng

RESEARCH ENGINEER AT MAUMAI

[✉ sanha.maeng@gmail.com](mailto:sanha.maeng@gmail.com) | [🏡 9rum.github.io](https://9rum.github.io) | [GitHub](https://github.com/9rum) | [LinkedIn](https://www.linkedin.com/in/9rum/) | [Google Scholar](https://scholar.google.com/citations?user=9rum)

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 Inc.

RESEARCH ENGINEER (TECHNICAL RESEARCH PERSONNEL)

- Improved LLM serving throughput by up to $352\times$ using continuous batching, PagedAttention and FlashAttention.
- 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. Further improved by $3\times$ with computation-I/O overlapping.

Seongnam, Republic of Korea

Mar. 2023 -

CONCAT Inc.

SOFTWARE ENGINEER INTERN

Seoul, Republic of Korea

Jun. 2020 - Sep. 2020

FlyHigh Co., Ltd.

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 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

KIISE

2020 **2nd Place**, KMUCS Capstone Design Conference

KMU