Seonggon Kim

Dept. of Computer Science and Engineering, POSTECH, Republic of Korea
sungonuni@postech.ac.kr

RESEARCH INTEREST

I'm currently focusing on Efficient AI, particularly in enhancing **memory efficiency** and **computation acceleration** during the **training** and **inference** of various models (Such as Vision, LLM, and video generation) via **quantization** and **low-rank approximation**.

KEYWORD

- Fast and Memory-Efficient Training *(A01, C03)*
- Fast and Memory-Efficient Inference *(C01, C02)*
- Parameter Efficient Fine-tuning of LLMs *(U01, P03)*
- CUDA Kernel optimization *(P01, P02, P03, C01, C03)*
- Fast Sampling of Video Generation Diffusion Models

EDUCATION

POSTECH Pohang, Korea

Ph.D. in Computer Science and Engineering Sep. 2023 – Present

Advised by Professor Eunhyeok Park.

KYUNG HEE UNIVERSITY Seoul, Korea

B.S. in Computer Science and Engineering Feb. 2017 - Aug. 2023

PUBLICATIONS

[C03] HOT: Hadamard-based Optimized Training
Seonggon Kim, Juncheol Shin, Seung-taek Woo, Eunhyeok Park
Computer Vision and Pattern Recognition (**CVPR 2025**), Nashville.

[C02] Merge-Friendly Post-Training Quantization for Multi-Target Domain Adaptation
Juncheol Shin, Minsang Seok, **Seonggon Kim**, Eunhyeok Park
International Conference on Machine Learning (**ICML 2025**), Vancouver.

[C01] PTQ4VM: Post-training Quantization for Visual Mamba
Younghyun Cho*, Changhun Lee*, **Seonggon Kim**, Eunhyeok Park
Winter Conference on Applications of Computer Vision (**WACV 2025 Oral**), Tucson.

[U01] HoLA: Overcoming the full-finetuning with Hadamard-oriented LoRA
Seonggon Kim, Taehyeon Kim, Byeori Kim, Eunhyeok Park
Neural Information Processing Systems (**NeurIPS 2025**, Under review), San Diego.

[A01] HLQ: Fast and Efficient Backpropagation via Hadamard Low-rank Quantization
Seonggon Kim, Eunhyeok Park
arXiv 2406.

PROJECT

[P03] Fast and Memory-efficient training on Extreme environment — Jul. 2024 – Current
National AI Research Lab of Korea — Seoul, Korea

- Conducted research on **memory-efficient training** for vision models.
- Prototype development of an optimized **CUDA kernel for memory-efficient training.**

[P02] GEMV Accelerator for LLM inference on Intel Gaudi-2 — Jun. 2024 - Jun. 2025
Naver & Intel Joint Research Center — Seoul, Korea

- Conducted research on **LLM's fast inference** on Intel Gaudi-2 architecture.
- Implemented custom **GEMV kernel for Gaudi** with TPC-C language.
- Transplanted LUT Quantization from CUDA to Gaudi TPC.

[P01] Solutions for Self-supervised training on Edge Device — Jun. 2023 – Current
Ministry of Science and ICT (Korean Government) — Daejeon, Korea

- Conducted research on **fast fine-tuning** on Edge device.
- Designed an efficient fine-tuning algorithm with stochastic quantization.
- Implemented custom **CUDA kernel for fast fine-tuning**.

EXPERIENCE

SOFTWARE ENGINEER INTERN — Jul. 2022 - Feb. 2023
Spirent Communications — San Jose, CA, USA

- C++ backend engineering on 5G testing frameworks 'LandSlide'

SOFTWARE ENGINEER INTERN — Feb. 2022 - Jun. 2022
Common Computer — Seoul, Korea

- Web3 Smart Contract engineering on Ethereum

RESEARCH INTERN — Mar. 2021 - Dec. 2021
SI Analytics — Daejeon, Korea

- Research on Semantic segmentation model for Satellite imagery
- Research on Unsupervised, Semi-supervised Learning and Domain Adaptation

AWARDS & HONORS

- ETHDenver 2022 Blockchain Hackathon, NFT project, 3rd Prize — Feb. 2022
- CVPR 2021 Earthvision workshop, Land Cover Classification Challenge, Selected as the final five teams — Jun. 2021

TEACHING EXPERIENCE

TEACHING ASSISTANT — Mar. 2025 - June. 2025

POSTECH — Pohang, Korea

- CSED311: Computer Architecture [2025-Spring]