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 Physical AI) via **quantization** and **low-rank approximation**.

KEYWORD

- Low-Precision Training *(A01, A02, C03, P01, P05)*
- Low-Precision Inference *(C01, C02, P02)*
- KV Cache Compression via Low-rank Approximation *(P04, PT01)*
- Parameter Efficient Fine-tuning of LLMs *(P03)*
- HW–SW Co-design via Kernel Optimization *(P01, P02, P04, P05)*

EDUCATION

POSTECH Pohang, Korea
M.S./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 Mar. 2017 - Aug. 2023

PUBLICATIONS

[A02] AdaHOP: Fast and Accurate Low-Precision Training via Outlier-Pattern-Aware Rotation
Seonggon Kim, Alireza Khodamoradi, Pranathi Vasireddy, Kristof Denolf, Eunhyeok Park
arXiv 2604.

[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.
*Equal contribution.

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

PATENTS

[PT01] Power Iteration Method Based on Hadamard PCA and KV Cache Compression Methodology Using the Same
Seonggon Kim, Taehyeon Kim, Eunhyeok Park
Korean Patent Application No. 10-2025-0189902, filed Dec. 03, 2025.

PROJECT

[P05] Low-Precision Training on AMD GPU architecture Nov. 2025 - May 2026
Advanced Micro Devices Longmont, CO, USA

- Conducted research on **Low-Precision (MXFP4) Training** on the AMD CDNA4 architecture.
- Implemented **ROCm kernel** for training acceleration.
- Achieved **3.6X memory compression, 1.46X acceleration** over BF16.
- Published in [A02].

[P04] Solutions for slow SVD approximation problem of KV Cache compression Feb. 2025 – Present
POSTECH Pohang, Korea

- Conducted research on the slow **SVD approximation** problem during KV Cache compression.
- Implemented **CUDA kernel** for SVD approximation acceleration.
- Achieved **5.1X acceleration and 95% similarity** with exact SVD operation.
- Filed as a **Korean patent** application [PT01].
- This work is currently under review.

[P03] Solutions for misaligned weight initialization problem of LoRA finetuning Jun. 2025 – Present

POSTECH Pohang, Korea

- Conducted research on **Effective LoRA weight initialization**.
- Achieved **99% accuracy of full fine-tuning** with only rank 16.
- This work is currently under review.

[P02] GEMV Accelerator for LLM inference on Intel Gaudi-2 Jun. 2024 - Jun. 2025

Naver & Intel Joint Research Center Seoul, Korea

- Conducted research on **fast LLM inference** on the 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 efficient training in limited GPU environments Jun. 2023 – Present

Ministry of Science and ICT of Korea Daejeon, Korea

- Conducted research on **Low-Precision (INT4) Training** in limited GPU environments.
- Implemented **CUDA kernel** for training acceleration.
- Achieved up to **75% memory savings and 2.6X acceleration** over FP32.
- Published in [A01], [C03].

EXPERIENCE

RESEARCH ASSOCIATE Nov. 2025 - May 2026

Advanced Micro Devices Longmont, CO, USA

- Research on Low-Precision (MXFP4) training and ROCm kernel optimization for the AMD CDNA4 architecture.

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

- Qualcomm Innovation Fellowship Korea, **Winner** Oct. 2025
- BK21 Outstanding Graduate Student International Training Scholarship, **Recipient** Oct. 2025
- CVPR 2021 EarthVision Workshop, Land Cover Classification Challenge, **5th Prize** Jun. 2021

TEACHING EXPERIENCE

TEACHING ASSISTANT Sep. 2026 - Dec. 2026
POSTECH Pohang, Korea
- CSED342: Artificial Intelligence [2026-Fall]

TEACHING ASSISTANT Mar. 2025 - Jun. 2025
POSTECH Pohang, Korea
- CSED311: Computer Architecture [2025-Spring]

ACADEMIC SERVICE

REVIEWER 2026
Conference on Neural Information Processing Systems (**NeurIPS 2026**)