

Research Interests

My interests lie in the field of Artificial Intelligence (AI) with a strong focus on deep learning systems and code optimization. I specialize in enhancing computational efficiency and improving model performance, with experience in tasks such as Game AI, Computer Vision, Natural Language Processing (NLP), and code parallelization/acceleration. Recently, I have been deeply involved in optimizing deep learning models, participating in AI competitions, and focusing on LLM quantization to push the boundaries of performance and efficiency in AI applications.

Education

- 2019- **Seoul National University**,
Ongoing *B.S. in Computer Science & Engineering, minor in Linguistics*, GPA 3.76/4.3.
Leave of absence for military service: Mar 2021 - Sep 2022

Research Experience

- 2023- **SNU ARC LAB**, *Internship*, (Advisor: Prof. Jae W. Lee).
Ongoing
 - Contributed to Any-Precision LLM experimentation and significantly enhanced the core quantization logic, achieving performance gains exceeding 10,000x.

2023 **SNU HCI LAB**, *UROP*, (Advisor: Prof. Jinwook Seo).
 - Worked on paper 'Accurate, Scalable, and Stable Dimensionality Reduction Using UMATO'
 - Optimized UMATO to reach performance comparable to SOTA DR techniques.
 - Helped create ZADU, a DR evaluating library (see publications)

Publications

- ICML Oral 2024 **Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLMs**,
Yeonhong Park, Jake Hyun, SangLyul Cho, Bonggeun Sim, Jae W. Lee, Oral Presentation (Top 1.5%).
 - Optimized core Any-Precision LLM quantization logic, and worked on model performance experimentations.
 - Implemented the official open-source library, enabling automatic application of our work on arbitrary LLMs.

IEEE VIS 2023 **ZADU: A Python Library for Evaluating the Reliability of Dimensionality Reduction Embeddings**,
Hyeon Jeon, Aeri Cho, Jinhwa Jang, Soohyun Lee, Jake Hyun, Hyung-Kwon Ko, Jaemin Jo, Jinwook Seo.
 - Converted GPU-accelerated evaluation code to optimize performance on CPU platforms.

Awards & Achievements

- 2024 **Accelerator Programming Winter School (CUDA competition)**, *1st place, team of two*,
[Organized by SNU THUNDER Research Group & Manycocoresoft].
 - 1st place by performance, final project on model inference throughput optimization using CUDA C++.

2022 **Korean AI Competition**, *1st place (Undergrad Div.), team of four, prize: KRW 10M*,
[Organized by Korea Ministry of Science and ICT, National Information Society Agency].
 - Developed a speech-to-text model for the Korean language & its dialects.
 - Awarded by Korean Minister of Science and Technology.

2020 **SNUH Medical AI Challenge**, *4th place, team of 11*,
[Organized by Seoul National University Hospital].
 - Developed an intraoperative hypotension predictor from arterial pressure waveforms.

2020 **Digital Health Hackathon**, *1st place, team of three, prize: KRW 3M*,
[Organized by Samsung Advanced Institute for Health Sciences & Technology, Digital Healthcare Partners].
 - Created a drug treatment decision model for rare cancer, based on a two-model ensemble approach.

2017 **Korean Olympiad in Informatics, project division**, *Silver(3rd place)*,
[Organized by Korea Ministry of Science and ICT].
 - Created an RL based AI agent for the game of Othello and Omok.

Open-Source Contributions

- 2024 **Any-Precision LLM**, [repo link](#),
[An LLM quantization library capable of quantizing and running variable bit-width models].
- Implements work from ICML paper of the same name (listed above).
 - Created the highly optimized yet versatile quantization pipeline.
- 2024 **flash1dkmeans**, [repo link](#),
[An optimized K-means implementation for the one-dimensional case].
- Devised, verified and implemented a variant of K-means clustering highly optimized for the 1D data.
 - Used directly in quantization works like Any-Precision LLM to bring down the quantization cost dramatically.
- 2023 **Steadiness & Cohesiveness**, [repo link](#),
[Quality metrics for evaluating the inter-cluster reliability of multidimensional projections].
- Parallelized the algorithm for distance matrix and Shared Nearest Neighbor (SNN) matrix calculations.
- 2023 **UMATO: Uniform Manifold Approximation with Two-phase Optimization**, [repo link](#),
[A dimensionality reduction technique that preserves both global and local structures of high-dimensional data].
- Reduced time complexity of core algorithm to reach performance comparable to UMAP, a SOTA technique.

Academic Project Highlights

- 2023 **LLVM Compiler Optimization Project**, *1st of 11 teams, Principles and Practices of Software Development*.
◦ Secured 1st place in competitive project optimizing compiler performance on custom system using LLVM passes.
- 2023 **Crowd-Analyzer**, *Creative Integrated Design*.
◦ Developed a CCTV crowd flow and density analysis system for security platform company INNODEP.
- 2022 **System Programming Lab Assignments**, *1st of 107, System Programming*.
◦ 1st by optimization score in tasks encompassing IO, dynamic memory, shell design and socket programming.
- 2020 **Computer Architecture Optimization Tasks**, *1st of 130, Computer Architecture*.
◦ 1st in all 4 competitive assignments, including floating point conversion and RISC-V assembly programming.

Personal Projects

- 2023 **Brick Breaker AI**, [repo link](#).
◦ Created a clone of a popular brick breaker game variant.
◦ Trained a neural network on self-generated labels, bootstrapping the process to create a competent AI agent.
- 2021 **Tranquil Tempest: An ISMCTS based Mighty AI**, [repo link](#).
◦ Personal project implementing an AI for the card game of Mighty based on ISMCTS, a variant of the MCTS algorithm for imperfect information games.

Extra Curriculars

- 2023 Working as website and server administrator for SNUPO(Seoul National University Philharmonic Orchestra).
- 2021-2022 Served as the Cybersecurity Team at the Cyber Defense Center, 7th Corps of ROK Army.
- 2020 Developed an SMS survey scheduler for Center for Happiness Studies, Seoul National University.
- 2019, 2020 UCPC(Union of Clubs for Programming Contests) programming contest finalist.
Participated in ACM-ICPC 2019 and 2020, Google Code Jam 2019 and 2020.
- 2018 ISEF-K(International Science and Engineering Fair Korea) participant, with improved Othello/Omok AI over the KOI competition(2017).

Skills

- Languages Python, C, C++
Frameworks PyTorch, Django, Numba
Technologies CUDA, LLVM Passes
Proficiencies Arduino, Raspberry Pi, Over 12 years of GNU/Linux experience.

Personal Details

- Language English & Korean, bilingual proficiency.