JANGHWAN LEE

222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea, 04763
superdocker.github.io superdocker +82-10-3220-0288 hwanii0288@hanyang.ac.kr

RESEARCH INTERESTS

Efficient deep learning inference/training algorithm; Post-training quantization; Reduced-precision numerical format; Transformer models; Large language models.

EDUCATION

Hanyang University, Seoul, Republic of Korea.
Integrated Ph.D. Student in Department of Electronic Engineering Mar. 2020 - Present
Artificial Intelligence Hardware & Algorithm Lab
Advisor: Professor Jungwook Choi

B.S. in Department of Electronic Engineering Mar. 2014 - Feb. 2020
Thesis: Fast face detector using DCT coefficients
Advisor: Professor Kiseok Chung

INTERNSHIP EXPERIENCE

Student Internship Program Jul. 2023 - Sep. 2023
Samsung Advanced Institute of Technology (SAIT), Suwon, Republic of Korea.
Research topic: Large-scale AI

RESEARCH EXPERIENCE

Research Assistant Mar 2020 - Present
Hanyang University Seoul, Republic of Korea
Advisor: Professor Jungwook Choi

- Understanding Microscaling Formats: Taming Activation Outliers through Asymmetry (Findings of ACL '25)
  - Analyzed why 4-bit microscaling (MX) formats are widely adopted by industry leaders (e.g., NVIDIA), particularly for mitigating activation outliers in LLM inference.
  - Found that microscaling addresses activation outliers effectively, but at the cost of increased data asymmetry due to small group size.
  - Proposed AMXFP4, an asymmetric 4-bit format that resolves this limitation, achieving better accuracy while maintaining hardware efficiency.
- Rank-Insensitive Error Compensation for 2-bit LLMs (AAAI '25)
  - Identified that quantization error in 2-bit LLMs is inherently high-rank, making direct weight-level compensation with low-rank LoRA ineffective.
  - Conducted the first analysis of the relationship between optimization granularity and compensation rank.
  - Discovered that at coarse granularity (e.g., model-wise), low-rank adapters become less sensitive to error rank.
  - Proposed RILQ, a model-wise activation discrepancy loss method that leverages this rank insensitivity to improve 2-bit quantization performance.
- Reduced-Precision LLM Inference Simulation Framework (2025)
  - Led development of `QLLM-INFER`, an open-source simulation framework for low-bit LLM inference in collaboration with Dnotitia Inc.
  - Built standardized evaluation pipelines on llama-3.1-8B for 8 quantization methods.

  - Quantified trade-offs in weight-activation, weight-only, and KV-cache quantization for real-world deployment.
  - Released on GitHub: `https://github.com/dnotitia/qllm-infer`
- Enhancing Conversational Ability in Quantized LLM-based Chatbots (ACL '24)
  - Identified "token-flipping" due to quantization as a major cause of degraded conversational ability of quantized LLMs.
  - Proposed QDPO, a quantization-aware preference optimization method that aligns outputs with full-precision models.
  - Restored conversational quality in both Korean and English chatbots, validated using GPT-4 evaluations.
- Development of Data Format for Weight-Activation Quantized LLM Inference (EMNLP '23)
  - Analyzed model-specific PTQ sensitivity in W4A8 inference across OPT and LLaMA.
  - Designed a denormal integer format to mitigate underflow from 4-bit weight quantization.
  - Achieved 2.56× power savings with negligible accuracy loss across 7B–30B models in hardware evaluations.
- Mathematical Analysis of Quantization Error in Fixed-Point and Floating-Point Arithmetic (ICASSP '23)
  - Modeled quantization errors in ViT operations and proposed a mixed-format algorithm for numerical format selection.
  - Developed a simple statistical test to guide format choice based on operation characteristics.
  - Achieved state-of-the-art accuracy with 6-bit PTQ for both weights and activations.
- Post-Training Quantization of Transformer Encoder Models with Sub-8-Bit Floating-Point (AICAS '22)
  - Optimized exponent bias in floating-point formats to reduce quantization error.
  - Introduced SQNR-based progressive bias tuning for post-training quantization.
  - Achieved near full-precision accuracy with 6–8-bit quantization on BERT for GLUE and SQuAD.
- Reduced-Precision Training Simulation Framework (2021)
  - Built a PyTorch CUDA-based simulation framework for low-bit training with configurable precision.
  - Enabled reduced precision for weights, activations, gradients, and accumulations.
  - Demonstrated no accuracy drop on SSD-Lite and ResNet/MobileNet with 8-bit training.

PUBLICATIONS

[**ACL 2025**] **Janghwan Lee**, Jiwoong Park, Jinseok Kim, Yongjik Kim, Jungju Oh, Jinwook Oh, and Jungwook Choi, "AMXFP4: Taming Activation Outliers with Asymmetric Microscaling Floating-Point for 4-bit LLM Inference", In Findings of the Association for Computational Linguistics (ACL Findings)

[**AAAI 2025**] Geonho Lee*, **Janghwan Lee***, Sukjin Hong*, Minsoo Kim, Euijai Ahn, Du-Seong Chang, and Jungwook Choi, "RILQ: Rank-Insensitive LoRA-based Quantization Error Compensation for Boosting 2-bit Large Language Model Accuracy", The 39th Annual AAAI Conference on Artificial Intelligence

[**ASAP 2024**] Dong-eon Won*, Yeeun Kim*, **Janghwan Lee**, Minjae Lee, Jonghyun Bae, Jongjoo Park, Jeongyong Song, and Jungwook Choi, "ISP2DLA: Automated Deep Learning Accelerator Design for On-Sensor Image Signal Processing", In 35th IEEE International Conference on Application-specific Systems, Architectures and Processors (Poster)

[**ACL 2024, Oral**] **Janghwan Lee***, Seongmin Park*, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi, "Improving Conversational Abilities of Quantized Large Language Models via Direct Preference Alignment", In The 62nd Annual Meeting of the Association for Computational Linguistics

[**HPCA 2024**] Minjae Lee, Seongmin Park, Hyungmin Kim, Minyong Yoon, **Janghwan Lee**, Junwon Choi, Nam Sung Kim, Mingu Kang, and Jungwook Choi, "SPADE: Sparse Pillar-based 3D Object Detection Accelerator for Autonomous Driving", 30th IEEE International Symposium on High-Performance Computer Architecture

[**ICEIC 2024, Oral**] Youngdeok Hwang*, **Janghwan Lee***, Jiwoong Park, Jieun Lim, and Jungwook Choi, "Searching Optimal Floating-Point Format for Sub-8-Bit Large Language Model Inference", In International Conference on Electronics, Information, and Communication

[**EMNLP 2023**] **Janghwan Lee***, Minsoo Kim*, Seungcheol Baek, Seok Joong Hwang, Wonyong Sung, and Jungwook Choi , "Enhancing Computation Efficiency in Large Language Models through Weight and Activation Quantization", The 2023 Conference on Empirical Methods in Natural Language Processing

[**NeurIPS 2023**] Minsoo Kim, Sihwa Lee, **Janghwan Lee**, Hong Sukjin, Chang Du-Seong, and Sung Won Yong, and Jungwook Choi, "Token-Scaled Logit Distillation for Ternary Weight Generative Language Models", Thirty-seventh Conference on Neural Information Processing System

[**ICASSP 2023**] **Janghwan Lee**, Youngdeok Hwang, and Jungwook Choi, "Finding Optimal Numerical Format for Sub-8-bit Post-Training Quantization of Vision Transformers", 2023 IEEE International Conference on Acoustics, Speech and Signal Processing

[**DAC 2023**] Janghyeon Kim, **Janghwan Lee**, JeongHo Han, Sangheon Lee and Jungwook Choi, "Range-Invariant Approximation of Non-Linear Operations for Efficiently Fine-tuning BERT", 60th ACM/IEEE Design Automation Conference

[**AICAS 2022, Oral**] **Janghwan Lee**, and Jungwook Choi, "Optimizing Exponent Bias for Sub-8bit Floating-Point Inference of Fine-tuned Transformers", 2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems

SCHOLARSHIP AND AWARD

Excellence in Research and Education, ENRICH IT AWARD, 4th Stage BK21 — Spring 2025
Integrated Ph.D. Course Scholarship, Full Tuition, Hanyang University — Spring 2020 - Fall 2022
Research Scholarship, Total KRW 24M, ISRC — Fall 2020 - Fall 2022
AI Grand Challenge, Korea Ministry of Science and ICT — Fall 2020

- First place award in Model Compression Track

B.S. Course Scholarship, Full Tuition, Hanyang University — Spring 2014 - Spring 2015, Fall 2018 - Fall 2019

SKILLS

Programming Languages	Python, C, C++
Deep Learning Frameworks	PyTorch, Huggingface
Systems & Platforms	Linux, Git, Docker