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 format. Floating-point. Transformer models. Large language models.

EDUCATION

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

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

INTERNSHIP EXPERIENCE

Student Internship Program Jul. 2023 - Sep. 2023
Samsung Advanced Institue of Thechnology (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

- Enhancing Conversational Ability in Quantized LLM-based Chatbots (ACL '24)
  - Analyze that the degradation in conversational ability of LLM-based chatbots in multi-turn conversations is due to *token-flipping*, where the Top-1 and Top-2 tokens differ as a result of quantization errors.
  - Propose a QDPO (Quantization-aware Direct Preference Optimization) method to restore the conversational ability of quantized LLMs by aligning them with their full-precision counterparts.
  - Improve conversational ability to baseline levels in both Korean and English language models as evaluated in chatbot conversation tasks using GPT-4.
- Development of Data Format for Weight-Activation Quantized LLM Inference (EMNLP '23)
  - Analyze the differences in model-specific data statistics and PTQ calibration sensitivity for W4A8 LLM inference on OPT and LLaMA models.
  - Develop a denormal integer format to prevent underflow by analyzing the impact of 4-bit weight quantization errors on output.
  - Confirm 2.56x power saving compared to conventional INT8 operations through hardware evaluation, with negligible degradation in W4A8 inference across 7B to 30B OPT and LLaMA models.
- Mathematical Analysis of Quantization Error in Fixed-Point and Floating-Point Arithmetic (ICASSP '23)
  - Analyze diverse data characteristics in Vision Transformer (ViT) operations and propose a *mixed-format* algorithm that optimizes numerical formats for each operation.
  - Develop a decision rule based on mathematical modeling of fixed-point and floating-point quantization errors, utilizing an efficient and simple statistical test.

  - Achieve state-of-the-art accuracy with post-training quantization of both weights and activations in ViT down to 6-bit precision.
- Post-Training Quantization of Transformer Encoder Models with Sub-8-Bit Floating-Point (AICAS '23)
  - Develop a practical optimization method for exponent bias in floating-point numbers, minimizing quantization errors.
  - Introduce SQNR (Signal to Quantization Noise Ratio) based progressive exponent bias optimization.
  - Attain near full-precision model accuracy with 6- to 8-bit floating-point post-training quantization of fine-tuned BERT on GLUE and SQuAD tasks.
- Reduced-Precision Training Simulation Framework
  - Implement reduced-precision training simulation framework on PyTorch's CUDA backend
  - Enable the adjustment of bit-widths for weights, activations, gradients, and partial-sum accumulation to simulate deep learning training on real-world hardware.
  - Demonstrate no performance degradation in object detection (SSD-Lite) and image classification models (ResNet18, ResNet50, and MobileNetV2) with 8-bit training.

PUBLICATIONS

[**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, **Janhwan 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, Dec 2023

[**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 (AICAS)

SCHOLARSHIP AND AWARD

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