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 Training. Floating-Point. Transformer Model. Large Language Model.

EDUCATION

Integrated Ph.D. Student in Department of Electronic Engineering Mar. 2020 - Present
Hanyang University, Seoul, Republic of Korea.

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

- Mathematical analysis of the quantization error of fixed-point and floating-point
  - Observe the diverse characteristics of data in the operation of the Vision Transformer(ViT), and propose a *mixed-format* algorithm optimizing the numerical formats for each operation in ViT
  - Decision rule based on mathematical modeling of fixed and floating-point quantization errors with efficient simple statistical test.
  - State-of-the-art accuracy with post-training quantization on both weights and activations in ViT to 6-bit
- Post-training quantization on Transformer encoder model with sub-8-bit floating-point
  - Practical optimization method for the exponent bias of floating-point minimizing quantization errors
  - SQNR(Signal to Quantization Noise Ratio) based progressive exponent bias optimization
  - Achieve close to full-precision model accuracy for 6 to 8 bit floating-point post-training quantization of fine-tuned BERT on GLUE and SQuAD tasks
- Reduced-precision training simulation framework
  - Implement PyTorch's CUDA backend enabling adjustment of the bit-widths of weight, activation, gradient, and partial-sum accumulation for simulation of deep learning training on real-world hardware
  - No performance degradation for object detection model(SSD-Lite), and image classification models(ResNet18, ResNet50, and MobileNetV2) with 8-bit training

PUBLICATIONS

[**HPCA 2024 Accept**] 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

[**EMNLP 2023 Accept**] **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 Poster**] 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 Poster**] **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 Poster**] 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)

Janghwan Lee, Sidong Roh, and Kiseok Chung, "Fast face detector using DCT coefficients", Korean Institute of Communications and Information Science Fall Conference 2019

SCHOLARSHIP AND AWARD

Integrated Ph.D. Course Scholarship, Full Tuition, Hanyang University — Spring 2020 - Spring 2023
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

SKILLS

Programming Languages Python, C, C++
Deep Learning Frameworks PyTorch, Huggingface