MINSOO KIM

+82-10-8203-6871 marsjacobs.github.io minsoo2333@hanyang.ac.kr

RESEARCH INTERESTS

General Efficiency for LLM/LMM Inference - quantization, kv-cache compression, long context LLM/LMM, multi-modality, self-improvement, knowledge distillation, parameter efficient fine-tuning, low-rank compression

EDUCATION

Hanyang University, Seoul, South Korea Mar. 2021 - Present
Artificial Intelligence Hardware & Algorithm lab
Ph.D. Student in Electronic Engineering
Advisor: Professor Jungwook Choi

Hanyang University, Seoul, South Korea Feb. 2021
B.S in Electronic Engineering
Thesis: Improving training method for very low bit weight quantization of Light Deep Learning Model
Advisor: Professor Jungwook Choi

WORK EXPERIENCE

Qualcomm AI Research, Seoul, South Korea, PhD research Intern Mar. 2024 - Present
Hanyang University, Seoul, South Korea, Student researcher Feb. 2021 - Present

PUBLICATIONS

- [**ACL 2024**] **Minsoo Kim**, Sihwa Lee, Wonyong Sung and Jungwook Choi "RA-LoRA: Rank-Adaptive Parameter-Efficient Fine-Tuning for Accurate 2-bit Quantized Large Language Models", *In Findings of the Association for Computational Linguistics: ACL 2024* (to appear)
- [**ACL 2024**] 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 Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (to appear)
- [**NeurIPS 2023**] **Minsoo Kim**, Sihwa Lee, Jangwhan Lee, Sukjin Hong, Du-Seong Chang, Wonyong Sung and Jungwook Choi "Token-Scaled Logit Distillation for Ternary Weight Generative Language Models", *Thirty-seventh Conference on Neural Information Processing Systems.* [Paper, Code]
- [**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", *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.* (*Co-First author) [Paper]
- [**EACL 2023**] **Minsoo Kim**, Kyuhong Shim, Seongmin Park, Wonyong Sung and Jungwook Choi, "Teacher Intervention: Improving Convergence of Quantization Aware Training for Ultra-Low Precision Transformers", *In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 916–929, Dubrovnik, Croatia. Association for Computational.* [Paper, Code
- [**EMNLP 2022**] **Minsoo Kim**, Sihwa Lee, Sukjin Hong, Du-Seong Chang, and Jungwook Choi, "Understanding and Improving Knowledge Distillation for Quantization-Aware Training of Large Transformer Encoders," *In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6713–6725, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.* [Paper, Code
- [**DAC 2022**] Joonsang Yu, Junki Park, Seongmin Park, **Minsoo Kim**, Sihwa Lee, Donghyun Lee, Jungwook Choi, "NN-LUT: neural approximation of non-linear operations for efficient transformer inference", *In Proceedings of the 59th ACM/IEEE Design Automation Conference.* [Paper]

RESEARCH EXPERIENCE

Research Intern, Qualcomm AI Research

- **Infinite Context (KV) Compression for Memory-Constrained LLMs - under review**
  - Iterative chunk-based context processing for memory-constrained recent LLMs (LLaMA/Mistral/Gemma/Phi)
  - Achieve 8x to 32x memory compression with comparable long context performance to GPT-3.5-turbo.
  - Analyze long context LLM characteristics in constrained memory environments - lost in the middle, retrieval

Research Assistance, Hanyang University (Advisor. Prof. Jungwook Choi)

- **Rank-Adaptive PEFT for 2-bit Quantized LLM LoRA Fine-Tuning - ACL 24**
  - Identify inherent high-rank property of low-bit LLM weight quantization error (LLaMA-2)
  - Investigate LoRA update behavior thorough singular value and vector analysis with SVD-based analysis
  - Propose rank adjusting method providing superior accuracy to SoTA quantized PEFT methods
- **Token-Scaled Logit Distillation (KD) for 2-bit (Ternary) LLMs - NeurIPS 23**
  - Quantization-Aware Training (QAT) on a generative language models (GPT-2/Neo, OPT, LLaMA)
  - Present confidence-based probabilistic correlation in the language modeling objective training
  - Propose novel KD method designed for LLM QAT, providing superior learning from teacher
- **LLMs 4-bit Weight and 8-bit Activation Quantization (PTQ) - EMNLP 23**
  - Analyze various LLM (OPT, LLaMA) characteristics of weight/activation distribution with quantization
  - Scaling & calibration PTQ method effectively addressing combined weight and activation quantization effects
  - Identify underflow in W4A8; propose hybrid data format and arithmetic unit with 2× HW efficiency
- **Improving KD for QAT of Large Transformer Encoders - EMNLP 22, EACL 23**
  - Analyze quantization effect on attention behavior in Transformer over various target NLU tasks
  - Improve accuracy in NLU for 2bit (ternary) weight quantization for BERT and RoBERTa
  - Achieve higher accuracy in BERT-base/large and ViT within up to 12.5x shorter fine-tuning time

HONORS AND AWARDS

- **AICAS Grand Challenge 2024**, SW&HW Co-Optimization for LLM, 3rd place — March 2024
- **Qualcomm Innovation Fellowship Korea 2023**, Winner, *Qualcomm* — November 2023
- **AI Grand Challenge**, 1st place, *Korea Ministry of Science and ICT* — November 2020
- **Reseach Scholarship** *IoT System Semiconductor Research Center* — Spring 2021 - Spring 2023

SKILLS

- **Programming Languages**: Python, C, C++
- **Teaching Assistant**: SOC design (Spring 2021), Introduction to SW Optimization (Fall 2023)
- **English**: Served as a KATUSA (Korean Augmentation to the US Army) (Jul 2017 - Apr 2019)
- **Academic Services**: Reviewer - NeurIPS, ICML, ACL, EMNLP, COLM, AAAI (2023 - present)