MINSOO KIM

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

RESEARCH INTERESTS

General Efficiency for LLM/LMM Inference - model quantization; long context LLM/LMM; multi-modality; knowledge distillation; parameter efficient fine-tuning

EDUCATION

Hanyang University, Seoul, South Korea Mar. 2021 - Feb. 2026
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

Apple, Seattle, ML Research Intern Mar. 2025 - Sep. 2025
Qualcomm AI Research, Seoul, Research Intern Mar. 2024 - Mar. 2025
Hanyang University, Seoul, Student Researcher Mar. 2021 - Present

SELECTED PUBLICATIONS

- [**EMNLP 2024**] **Minsoo Kim**, Kyuhong Shim, Jungwook Choi, and Simyung Chang, "InfiniPot: Infinite Context Processing on Memory-Constrained LLMs", *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* [Paper]
- [**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* [Paper]
- [**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* [Paper]
- [**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

RESEARCH EXPERIENCE

Research Intern, Qualcomm AI Research

- **Infinite Context (KV-Cache) Compression for Memory-Constrained LLMs - EMNLP 24**
  - Iterative chunk-based context processing for memory-constrained LLMs (LLaMA-3/Mistral/Gemma-2)
  - Achieve 8x to 32x memory compression with comparable long context performance to GPT-3.5-turbo
  - Propose novel token importance score perspective (past and future perspective) for KV-cache compression

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

- **Rank-Adaptive PEFT (LoRA) for 2-bit Quantized LLM Fine-Tuning - ACL 24**
  - Identify inherent high-rank property of low-bit LLM weight quantization error (LLaMA-2)
  - Investigate LoRA update behavior through 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, 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 impacts on self-attention mechanism in Transformer over NLU(GLUE) tasks
  - Propose attention KD for 2bit (ternary) weight-only quantization for BERT and RoBERTa (EMNLP 22)
  - Propose novel KD technique in BERT and ViT for speed-up fine-tuning time up to 12.5x (EACL 23)

HONORS AND AWARDS

- **Outstanding Reviewer** EMNLP 24 — November 2024
- **AICAS Grand Challenge 2024**, SW&HW Co-Optimization for LLM, 3rd place — March 2024
- **Qualcomm Innovation Fellowship Korea 2023**, Winner, Qualcomm AI Research — 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, ICLR, ICML, AISTATS, COLM, AAAI, ACL(ARR) (2023 - present)
  - Student Volunteer - EMNLP 22, 23, 24