MINSOO KIM

222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea, 04763
☎ +82-10-8203-6871 ⌂ marsjacobs.github.io ✉ minsoo2333@hanyang.ac.kr

RESEARCH INTERESTS

Efficient Deep Learning Algorithm
- Model quantization, Knowledge Distillation
Large Language Model Fine-Tuning
- Domain specific fine-tuning, parameter-efficient fine-tuning
Interpretability of Language Model
- Interpreting and analyzing the behavior of NLP models

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

PUBLICATIONS

- **[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 main]** 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 main] 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 Poster]
- **[EMNLP 2022 main] 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, Poster]
- **[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 Assistant, AI Algorithm & Hardware Lab, Hanyang University — Mar 2021 - Present
Advisor: Professor Jungwook Choi — *Seoul, South Korea*

- **Token-Scaled Logit Distillation for Ternary Weight Generative Language Models**
  - Investigate the challenges of applying Quantization-Aware Training on a decoder-based Transformer model
  - Identify and analyze cumulative quantization error observed in causal attention of decoder model
  - Present confidence-based probabilistic correlation in the training objective of language modeling
  - Propose novel KD method designed for GLM (up to 7B-sized) QAT, providing superior learning from teacher
- **Enhancing Efficiency in LLMs via Weight and Activation Quantization**
  - Analyze various LLM (OPT, LLaMA) characteristics of weight/activation distribution in quantization
  - Scaling & calibration based PTQ method addressing combined weight and activation effects on LLM
  - Identify underflow in W4A8; propose hybrid data format and arithmetic unit with 2× HW efficiency
- **Understanding and Improving KD for QAT of Large Transformer Encoders**
  - Mechanism of KD conducting attention recovery of quantized large Transformer encoders
  - Analyze quantization effect on attention behavior in Transformer over various target NLU tasks
  - Improve accuracy drop in NLU for 2bit weight quantization for large Transformer encoder with <1%
- **Improving Convergence of Quantization Aware Training for Ultra-Low Precision Transformers**
  - Proactive Teacher Intervention KD method for fast converging 2-bit QAT of Transformer encoders
  - Gradual Intervention Mechanism stabilizing the recovery of subsections of quantized Transformer layers
  - Achieve higher accuracy in BERT and ViT within up to 12.5x shorter fine-tuning time

SCHOLARSHIP AND AWARD

- **Qualcomm Innovation Fellowship Korea 2023**, Winner, USD 3K, *Qualcomm* — November 2023
- **Reseach Scholarship** USD 16K, *IoT System Semiconductor Research Center* — Spring 2021 - Spring 2023
- **AI Grand Challenge**, *Korea Ministry of Science and ICT* — November 2020
  - First place award in Model Compression Track (YOLOV5s Object Detection model 4x speed up)

SKILLS

- **Programming Languages**: Python, C, C++
- **Teaching Assistant**: SOC design (Spring 2021), Introduction to SW Optimization (Fall 2023)
- **DL Frameworks**: Pytorch, Huggingface
- **Cloud Computing Platform**: NAVER NSML Machine Learning platform, KT Genie Mars Server Platform
- **English**: Served as a KATUSA (Korean Augmentation to the US Army) and certified Air Traffic Control Operator of the 8th Army (Jul 2017 - Apr 2019)

ACADEMIC SERVICES

- Reviewer of EMNLP 2023