

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 Inference Algorithm, Model Quantization, Knowledge Distillation, Large Language Model

EDUCATION

Ph.D. Candidate in Department of Electronic Engineering

Mar. 2021 - Present

Hanyang University, Seoul, South Korea

B.S in Department of Electronic Engineering

Feb. 2021

Hanyang University, Seoul, South Korea

Thesis: Improving training method for very low bit weight quantization of Light Deep Learning Model

Advisor: Professor Jungwook Choi

RESEARCH EXPERIENCE

Research Assistant, Hanyang University

Mar 2021 - Present

Advisor: Professor Jungwook Choi

Seoul, South Korea

- **GPT based generative LLM compression & auto-regressive text generation operation analysis**
 - Analyze the biased word generation behavior in GPT-2 models under 2-bit weight quantization with knowledge distillation.
 - Propose new scaled KD method emphasizing low confidence word prediction region in FP teacher model, achieving comparable perplexity to FP teacher model with 2-bit weight quantized gpt-2 model.
 - Profiling the text generation inference operation in single GPU for GPT-2/3 models, identifying memory-bound and low-density computation challenges in GPU architecture with text-generation tasks.
- **Large Transformer encoder model QAT with Knowledge Distillation**
 - In-depth analysis of the mechanism of KD on attention recovery of quantized large Transformer encoders.
 - Analyze quantization effect on attention behavior of transformer over various language understanding tasks.
 - Propose a new KD method and unification of multiple KD loss function to address task-dependent preference.
 - Achieve state-of-the-art language understanding accuracy for QAT with sub-2bit weight quantization for large Transformer encoder models.
- **Improving Transformer encoder QAT convergence & accuracy of few-sample fine-tuning**
 - Propose a proactive Teacher Intervention KD method for fast converging QAT of low precision pre-trained Transformers.
 - Gradual intervention mechanism to stabilize the recovery of subsections of Transformer layers from quantization.
 - Achieves higher accuracy of language understanding task within 12.5x shorter fine-tuning time.

PUBLICATIONS

- **[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", *The 17th Conference of the European Chapter of the Association for Computational Linguistics* [Paper, Code]
- **[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," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* [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," *Proceedings of the 59th ACM/IEEE Design Automation Conference* [Paper]
- Hyeonseung Kim, **Minsoo Kim**, Jungwook Choi, "Improving training method for very low bit weight quantization of Light Deep Learning Model," Autumn Annual Conference of IEIE 2020

SCHOLARSHIP AND AWARD

- **Integrated PhD Course Scholarship**, Full Tuition, *Hanyang University* Spring 2021 - Spring 2024
- **Research Scholarship** USD 8K per year, *ISRC* Spring 2021 - Spring 2023
- **AI Grand Challenge**, *Korea Ministry of Science and ICT* Fall 2020
 - First place award in Model Compression Track
 - compress YOLOV5s Object Detection model with 4x speed up

SKILLS

- **Programming Languages:** Python, C, C++
- **DL Frameworks:** Pytorch, Huggingface
- **Cloud Computing Platform:** NAVER NSML Machine Learning platform, KT Genie Mars Server Platform
- **English Skill:** TOEIC 955, Served military service as KATUSA (Korean augmented to the US Army) in 8th Army (Sep 2017 - Apr 2019)