MINSOO KIM

222, Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea, 04763
+82-10-8203-6871 minsoo2333@hanyang.ac.kr https://github.com/MarsJacobs

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*

- **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 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 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.

Undergraduate Research Intern, Hanyang University Jul 2020 - Feb 2021
Advisor: Professor Jungwook Choi *Seoul, South Korea*

- **Fine-Tuning scheduling method for 2-bit weight quantization of light deep learning models**
  - Propose a better training scheduling method for boosting quantized model accuracy.
  - Improve 2-bit weight quantization accuracy of light deep learning models including EfficientNetB0 and MobileNetV2.

PUBLICATIONS

- **Minsoo Kim**, Kyuhong Shim, Seongmin Park, Wonyong Sung and Jungwook Choi, "Teacher Intervention: Improving Convergence of Quantization Aware Training for Ultra-Low Precision Transformers", **EACL 2023**
- **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," **EMNLP 2022** Paper, Code
- 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, " **DAC 2022**
- 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

AWARD

2020 AI Grand Challenge *Korea Ministry of Science and ICT*

- 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)