SEHOON KIM

1929 Delaware, Berkeley, CA 94709

✉ sehoonkim@berkeley.edu 510-960-9631 sehoonkim.org Google Scholar GitHub

RESEARCH INTERESTS

Efficient Deep Learning, Model Compression, Hardware-Software Co-design, AI Systems

EDUCATION

University of California at Berkeley Aug. 2020 - Present
Berkeley Artificial Intelligence Research (BAIR)
Ph.D. student in Electrical Engineering and Computer Science

Seoul National University Mar. 2015 - Feb. 2020
B.S. in Electrical and Computer Engineering
GPA: Overall **4.29/4.30**, Major **4.30/4.30**, Ranked **1st** in the entire class of 2020

Korea Science Academy of KAIST Mar. 2011 - Feb. 2015
Math and science specialized high school

PUBLICATIONS

- **Sehoon Kim***, Amir Gholami*, Albert Shaw†, Nicholas Lee†, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, Kurt Keutzer, "Squeezeformer: An Efficient Transformer for Automatic Speech Recognition," NeurIPS 2022 [Paper] [Code]
- Woosuk Kwon*, **Sehoon Kim***, Michael W. Mahoney, Joseph Hassoun, Kurt Keutzer, Amir Gholami, "A Fast Post-Training Pruning Framework for Transformers," NeurIPS 2022 [Paper] [Code]
- **Sehoon Kim***, Sheng Shen*, David Thorsley*, Amir Gholami*, Woosuk Kwon, Joseph Hassoun, Kurt Keutzer, "Learned Token Pruning for Transformers," KDD 2022 [Paper] [Code]
- **Sehoon Kim**, Amir Gholami, Zhewei Yao, Nicholas Lee, Patrick Wang, Anirudda Nrusimha, Bohan Zhai, Tianren Gao, Michael W. Mahoney, Kurt Keutzer, "Integer-only Zero-shot Quantization for Efficient Speech Recognition," ICASSP 2022 [Paper] [Code]
- Shixing Yu*, Zhewei Yao*, Amir Gholami*, Zhen Dong*, **Sehoon Kim**, Michael W Mahoney, Kurt Keutzer, "Hessian-Aware Pruning and Optimal Neural Implant," WACV 2022 [Paper]
- Gyeong-In Yu, Saeed Amizadeh, **Sehoon Kim**, Artidoro Pagnoni, Ce Zhang, Byung-Gon Chun, Markus Weimer, Matteo Interlandi, "WindTunnel: Towards Differentiable ML Pipelines Beyond a Single Model," VLDB 2022 [Paper]
- Taebum Kim, Eunji Jeong, Geon-Woo Kim, Yunmo Koo, **Sehoon Kim**, Gyeong-In Yu, Byung-Gon Chun, "Terra: Imperative-Symbolic Co-Execution of Imperative Deep Learning Programs," NeurIPS 2021
- **Sehoon Kim***, Amir Gholami*, Zhewei Yao*, Michael W. Mahoney, Kurt Keutzer, "I-BERT: Integer-only BERT Quantization," ICML 2021 (**Oral**) [Paper] [Code1] [Code2]

PREPRINTS and BOOK CHAPTERS

- **Sehoon Kim***, Coleman Hooper*, Thanakul Wattanawong, Minwoo Kang, Ruohan Yan, Hasan Genc, Grace Dinh, Qijing Huang, Kurt Keutzer, Michael W. Mahoney, Yakun Sophia Shao, Amir Gholami, "Full Stack Optimization of Transformer Inference: a Survey," Preprint 2023
- **Sehoon Kim***, Karttikeya Mangalam, Jitendra Malik, Michael W. Mahoney, Amir Gholami, Kurt Keutzer, "Big Little Transformer Decoder," Preprint 2023 [Paper] [Code]
- Amir Gholami*, **Sehoon Kim***, Zhen Dong*, Zhewei Yao*, Michael W. Mahoney, Kurt Keutzer, "A Survey of Quantization Methods for Efficient Neural Network Inference," Book Chapter: Low-Power Computer Vision: Improving the Efficiency of Artificial Intelligence, 2021

RESEARCH EXPERIENCES

Research Assistance, UC Berkeley — Aug. 2020 - Present
Advisor: Prof. Kurt Keutzer

- **Squeezeformer: An Efficient Transformer for Automatic Speech Recognition**
  - A next-generation attention-convolution hybrid architecture for efficient Automatic Speech Recognition
  - Temporal U-Net structure, which reduces the cost of the multi-head attention on long sequences, along with careful design of macro and micro-architecture
  - Achieved up to 3.1% word-error-rate reduction on LibriSpeech benchmark compared to state-of-the-art Conformer model under same FLOPs constraint
- **Learned Token Pruning for Transformers**
  - Token pruning scheme for Transformers that detects and drops less important tokens for efficient inference
  - Proposed fully-automated algorithm for determining optimal token pruning configuration by introducing learnable binary mask for tokens
  - Achieved 2.1× FLOPs reduction and up to 2× throughput improvement on Haswell CPU and V100 GPU with less than 1% accuracy degradation from RoBERTa
- **Integer-only Zero-shot Quantization for Efficient Speech Recognition**
  - Integer-only quantization scheme for ASR models that does not require any training/validation data
  - Proposed synthetic data generation method for speech signals that allows accurate calibration for quantization
  - Implemented on top of various ASR models and achieved 2.35× speedup of T4 GPU with less than 1% word-error-rate degradation
- **I-BERT: Integer-only BERT Quantization**
  - Integer-only quantization scheme for Transformers that performs entire inference with integer arithmetic
  - Introduced efficient and accurate integer-only kernels for GELU, Softmax, and LayerNorm, based on approximation with 2nd-order polynomials
  - Implemented I-BERT on top of RoBERTa and achieved 4× speedup on T4 GPU compared to FP32 baseline without accuracy degradation on GLUE benchmarks
  - **Open-source Project:** Collaborated with HuggingFace team to support I-BERT in official library

HONORS and AWARDS

Doctoral Study Abroad Scholarship, *Korea Foundation for Advanced Studies* — Up to five years from 2020
Full tuition, insurance, and living expenses (around 40 students selected nationally)

Kwanjeong Educational Foundation Scholarship, USD 10K per year — Spring 2017 - Fall 2018

Eminence Scholarship, Full Tuition, *Seoul National University* — Spring 2016 - Fall 2016

The Education and Research Foundation Scholarship, Full Tuition, *Seoul National University* — Fall 2015

Merit-based Scholarship, 10% Tuition, *Seoul National University* — Spring 2015

SKILLS

Programming Languages	Python, C/C++, JavaScript
AI Frameworks	PyTorch, Tensorflow, Keras
HW Simulation Tools	GEM5, CACTI
English Skill	iBT: 114 (R29, L30, S26, W29), GRE: Verbal 158, Writing 4.5