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. candidate 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

SELECTED PUBLICATIONS

- **Sehoon Kim***, Coleman Hooper*, Amir Gholami*, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, Kurt Keutzer, "SqueezeLLM: Dense-and-Sparse Quantization," Preprint 2023 [Paper] [Code]
- **Sehoon Kim**, Karttikeya Mangalam, Suhong Moon, John Canny, Jitendra Malik, Michael W. Mahoney, Amir Gholami, Kurt Keutzer, "Speculative Decoding with Big Little Decoder," Preprint 2023 [Paper] [Code]
- **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]

SURVEYS 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 (Short Version at ISCA ASSYST Workshop 2023) [Paper]

- Amir Gholami, Zhewei Yao, **Sehoon Kim**, Michael W. Mahoney, Kurt Keutzer, "AI and Memory Wall," RISELab Medium Blogpost, 2021 [Post]
- 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 [Paper]

RESEARCH EXPERIENCES

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

- **SqueezeLLM: Dense-and-Sparse Quantization**
  - Novel sensitivity-based non-uniform quantization scheme for LLMs that allocates quantization bins to more sensitive weight values to minimize post-quantization performance degradation
  - Dense-and-Sparse decomposition that isolates outliers in sparse matrix for better quantization performance
  - Lossless 4-bit and near-loseless 3-bit quantization of various LLMs with 2.3× latency improvement
- **Speculative Decoding with Big Little Decoder**
  - Collaborative use of small and large models where smaller model runs to autoregressively generates tokens and larger model reviews when challenging vocabularies appear
  - Simple fallback/rollback policies deciding when to use large model and when to reject small model's predictions
  - Up to 2× speedup on T4 GPU with minimal quality degradation on various generative tasks
- **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 sequence lengths for reduced inference costs, along with careful redesign of macro and micro-architecture
  - Up to 3% word-error-rate reduction on LibriSpeech compared to state-of-the-art Conformer with same FLOPs
- **Learned Token Pruning for Transformers**
  - Token pruning scheme for Transformers that detects and drops less important tokens for efficient inference
  - Fully-automated algorithm for determining optimal token pruning configuration by learnable binary mask
  - 2× FLOPs reduction and throughput improvement on Haswell CPU and V100 GPU with <1% accuracy drop
- **I-BERT: Integer-only BERT Quantization**
  - Integer-only quantization scheme for Transformers that performs entire inference with integer arithmetic
  - Integer-only kernels for non-linear operations through accurate approximation using 2nd-order polynomials
  - 4× speedup on T4 GPU compared to FP32 baseline without accuracy degradation on GLUE benchmarks
  - **Open-source Project:** Collaborated with HuggingFace 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