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

WORK EXPERIENCE

Narada AI, ML and Software Engineer May. 2022 - Present
University of California at Berkeley, Graduate Student Researcher Aug. 2020 - Present

SELECTED PUBLICATIONS

- **Sehoon Kim***, Suhong Moon*, Ryan Tabrizi, Nicholas Lee, Michael W. Mahoney, Kurt Keutzer, Amir Gholami, "An LLM Compiler for Parallel Function Calling," ICML 2024 [Paper] [LlamaIndex] [LangChain] [Code]
- **Sehoon Kim***, Coleman Hooper*, Amir Gholami*, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, Kurt Keutzer, "SqueezeLLM: Dense-and-Sparse Quantization," ICML 2024 [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," NeurIPS 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] [NVIDIA Nemo] [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]
- 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] [HuggingFace] [Code]

SURVEYS and BOOK CHAPTERS

- Amir Gholami, Zhewei Yao, **Sehoon Kim**, Michael W. Mahoney, Kurt Keutzer, "AI and Memory Wall," IEEE MICRO Journal Special Issue, 2024 [Paper] [Blog Post]
- **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 (ISCA ASSYST Workshop 2023) [Paper]
- 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]

HONORS and AWARDS

MLCommons ML and Systems Rising Star May. 2024
41 Ph.D. students worldwide in the fields of ML Systems

NVIDIA Graduate Fellowship Program Finalist Dec. 2023
15 Ph.D. students worldwide in the fields of computing innovation

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

RESEARCH EXPERIENCES

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

- **An LLM Compiler for Parallel Function Calling**
  - Framework for efficient and accurate LLM applications that enables optimized and parallel orchestration for multiple function calls with both open-source and closed-source models.
  - Up to 4× speedup, 7× cost savings, and 9% accuracy improvement compared to ReAct on various benchmarks.
  - **Official LlamaIndex** [Link] **and LangChain** [Link] **Integrations**. +1k stars on GitHub [Link]
- **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**
  - 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
  - **Official NVIDIA Nemo Library Integration** [Link]
- **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
  - **Official HuggingFace Library Integration** [Link]