SEHOON KIM

1929 Delaware, Berkeley, CA 94709
✉ sehoonkim@berkeley.edu ☎ 510-960-9631 github.com/kssteven418

RESEARCH INTERESTS

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

EDUCATION

University of California at Berkeley Aug. 2020 - Present
Candidate for *Ph.D. 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 class of 2020

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

PUBLICATIONS and PREPRINTS

- Woosuk Kwon*, **Sehoon Kim***, Michael W. Mahoney, Joseph Hassoun, Kurt Keutzer, Amir Gholami, "A Fast Post-Training Pruning Framework for Transformers," Preprint (Under Review) [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]

RESEARCH EXPERIENCES

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

- **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

Undergraduate Research Intern, Software Platform Lab, SNU Mar. 2019 - May. 2020
Advisor: Prof. Byung-Gon Chun

- **Terra: Imperative-Symbolic Co-Execution of Imperative Deep Learning Programs**
    - Framework that co-executes imperative DL programs and their optimized symbolic graph representations to achieve both flexibility of imperative programs and high-performance of symbolic programs.
- **WindTunnel: Towards Differentiable ML Pipelines Beyond a Single Model**
    - Framework that translates pre-trained classical machine learning models into equivalent neural networks to apply backpropagation for further improvement of model accuracy

Undergraduate Research Intern, High Performance Computer System Lab, SNU Sep. 2017 - Jun. 2018
Advisor: Prof. Jangwoo Kim

- **Power and Delay Simulator for SRAM at Ultra-low Temperature**
    - Tool that simulates delay, static power and dynamic power of SRAM architectures based on theoretically modeled physical characteristics of CMOS devices and wires at 77 K

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++, Verilog, Java, MATLAB
DL Frameworks	PyTorch, Tensorflow, Keras
HW Simulation Tools	GEM5, CACTI
English Skill	iBT: 114 (R29, L30, S26, W29), GRE: Verbal 158, Writing 4.5