Jiecao Yu

Contact Information

Website: https://jiecaoyu.github.io/ (Google Scholar)
E-mail: jiecaoyu@fb.com
Tel: +1(734)353-8285

Research Interests

Computer Architecture, Software-Hardware Co-design for Deep Learning Acceleration, DNN Pruning and Quantization.

Education

Ph.D. Candidate, Computer Science & Engineering 08/2014-09/2019
Advisor: Prof. Scott Mahlke
University of Michigan, Ann Arbor, MI
Dissertation: Efficient Deep Neural Network Computation on Processors

M.S. Computer Science & Engineering 08/2014-12/2015
University of Michigan, Ann Arbor, MI
Cumulative GPA: 4.00/4

B.Eng. Electronic & Information Engineering 08/2010-06/2014
Honored Minor, Advanced Honor Class of Engineering Education (ACEE)
Zhejiang University, Hangzhou, China
Cumulative GPA: 92/100 (3.98/4.0), Rank: 2/92

Publications

[1] Michael Anderson *et al.*, Jiecao Yu. “*First-Generation Inference Accelerator Deployment at Facebook*”. preprint at arXiv: 2107.04140

[2] Xiaowei Wang, Vidushi Goyal, Jiecao Yu, Valeria Bertacco, Andrew Boutros, Eriko Nurvitadhi, Charles Augustine, Ravi Iyer and Reetuparna Das. “*Compute-Capable Block RAMs for Efficient Deep Learning Acceleration on FPGAs*”. The 29th IEEE International Symposium On Field-Programmable Custom Computing Machines (FCCM-29), May, 2021

[3] Xiaocong Du, Bhargav Bhushanam, Jiecao Yu, Dhruv Choudhary, Tianxiang Gao, Sherman Wong, Louis Feng, Jongsoo Park, Yu Cao, Arun Kejariwal. “*Alternate Model Growth and Pruning for Efficient Training of Recommendation Systems*”. The 20th IEEE International Conference on Machine Learning and Applications (ICMLA-20), Dec, 2021

[4] Mao Ye, Dhruv Choudhary, Jiecao Yu, Ellie Wen, Zeliang Chen, Jiyan Yang, Jongsoo Park, Qiang Liu, Arun Kejariwal. “*Adaptive Dense-to-Sparse Paradigm for Pruning Online Recommendation System with Non-Stationary Data*”. preprint at arXiv: 2010.08655

[5] Jiecao Yu, Andrew Lukefahr, Reetuparna Das, Scott Mahlke. “*TF-Net: Deploying Sub-Byte Deep Neural Networks on Microcontrollers*”. ESWEEK-TECS special issue / the International Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES), Oct, 2019

[6] Jiecao Yu, Jongsoo Park, Maxim Naumov. “*Spatial-Winograd Pruning Enabling Sparse Winograd Convolution*”. preprint at arXiv: 1901.02132

[7] Xiaowei Wang, Jiecao Yu, Charles Augustine, Ravi Iyer, Reetuparna Das. “*Bit*

Prudent In-Cache Acceleration of Deep Convolutional Neural Networks". The 25th International Symposium on High-Performance Computer Architecture (HPCA-25), Feb, 2019

[8] Jiecao Yu, Andrew Lukefahr, David Palframa, Ganesh Dasika, Reetuparna Das, Scott Mahlke. "*Scalpel: Customizing DNN Pruning to the Underlying Hardware Parallelism*". The 44th International Symposium on Computer Architecture (ISCA-44), Jun, 2017

[9] Jiecao Yu, Andrew Lukefahr, Shruti Padmanabha, Reetuparna Das, Scott Mahlke. "*Adaptive Cache Partitioning on a Composite Core*". The PRISM-3 Workshop at the International Symposium on Computer Architecture (ISCA-42), Jun, 2015

Experiences

Facebook, Inc. (Meta Platforms, Inc.) 10/2019-Present
Senior Research Scientist (08/2021-Present) *Menlo Park, CA*
Research Scientist (10/2019-08/2021)
AI System SW/HW Co-design Group, Infrastructure

- Optimizing performance of ads models on Intel accelerators.
- Working on DNN model pruning and acceleration.

University of Michigan 08/2014-09/2019
Graduate Student Research Assistant *Ann Arbor, MI*

- Investigating the training algorithms of binary/ ternary neural networks.
- Developing low-precision computation algorithms/ hardware architecture for mobile and embedded devices.
- Developed a new DNN pruning technique, Scalpel, which applies weight pruning and node pruning synergistically based on the underlying hardware platform to improve the computation performance.

Facebook, Inc. 05/2018-08/2018
Research Intern, AI System SW/HW Co-design Group *Menlo Park, CA*
Manager: Dr. Jongsoo Park

- Proposed a two-step pruning technique, spatial-Wingorad pruning, to improve the Winograd-domain sparsity.

Arm, Inc. 05/2017-07/2017
Research Intern, Machine Learning Group *Austin, TX*
Manager: Dr. Ganesh Dasika

- Profiling and analysis of image captioning workloads (Show-and-Tell/Show-Attend-and-Tell).
- Built and profiled the server-side image captioning/classifying workloads based on TensorFlow Serving.

Arm, Inc. 06/2016-08/2016
Research Intern, Machine Learning Group *Austin, TX*
Manager: Dr. David Palframan

- Worked on Deep Neural Network acceleration on Arm cores, especially low-power microcontrollers.
- DNN weight pruning techniques are employed to compress the DNN in the keyword spotting (KWS) system.
- Libraries for sparse matrix computation on Arm Cortex-M4 microcontrollers are implemented and well-optimized.

University of Southern California 07/2013-09/2013

Research Intern *Los Angles, CA*
Supervisor: Prof. Melvin Breuer

- Worked on enhancing yield of VLSI chips via redundancy.

Patents

US 20180373975, ”Systems and Devices for Compressing Neural Network Parameters”, Jiecao Yu, Andrew Lukefahr, David Palframan, Ganesh Dasika, Reetuparnda Das, Scott Mahlke, Filed: June 21, 2017

US 20180373978, ”Systems and Devices for Formatting Neural Network Parameters”, Jiecao Yu, Andrew Lukefahr, David Palframan, Ganesh Dasika, Reetuparnda Das, Scott Mahlke, Filed: June 21, 2017

US 20170262285, ”Controlling Transition Between Using First and Second Processing Circuitry”, Andrew Lukefahr, Shruti Padmanabha, Jiecao Yu, Reetuparna Das, and Scott Mahlke, Filed: March 08, 2016

Talks & Posters

[Talk] Jiecao Yu. “*Efficient Low-Precision Deep Neural Networks on IoT Microcontrollers*”. Arm Research Summit, Sep, 2019

[Poster] Babak Zamirai, Jiecao Yu, Salar Latifi, Scott Mahlke. “*Input-specialized Heterogeneous Neural Networks*”. C-FAR 2016 Annual Meeting, Dec, 2016

[Poster] Salar Latifi, Babak Zamirai, Jiecao Yu, Scott Mahlke. “*Quality Assurance for Approximate Computing*”. C-FAR 2016 Annual Meeting, Dec, 2016

[Poster] Jiecao Yu, Babak Zamirai, Scott Mahlke. “*An Interactive Deep Neural Network Pruning System*”. C-FAR 2016 Semi-Annual Meeting, May, 2016

Service

Reviewer:

- Design Automation Conference (DAC) Technical Program Committee (TPC) Member ('20, 21, 22)
- IEEE Transactions on Computers ('20, 21)
- IEEE Access ('19, 20, 21)
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS'19, 20, 21)
- Elsevier Journal of Systems Architecture (JSA'19)
- ACM Journal on Emerging Technologies in Computing Systems (JETC'17)

Second Reviewer:

- Int. Conference on Compilers, Architecture, and Synthesis for Embedded Systems (CASES'17, 18, 19)
- Int. Symposium on Microarchitecture (MICRO'17, 19)
- Int. Symposium on Computer Architecture (ISCA'15, 17)
- Int. Symposium on Code Generation and Optimization (CGO'16, 17)
- Int. Symposium on High-Performance Computer Architecture (HPCA'16, 17)
- Int. Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'17)
- Int. Conference on Supercomputing (ICS'16)

Course Projects

L1 Cache Partitioning on a SMT Core Winter 2015
Parallel Computer Architecture (EECS 570), Prof. Thomas Wenisch
Designed a dynamic L1 cache partitioning mechanism for a SMT core based on way-partitioning technique and augmented LRU replacement policy. Cache capacities can

be resized at a fine granularity to capture the change of the cache demands for different threads.

A Two-Way Superscalar R10K SMT Processor Fall 2014
Computer Architecture (EECS 470), Prof. Trevor Mudge
Designed and implemented a synthesizable two-way superscalar Out-of-Order processor in Verilog HDL with speculative LSQ, instruction prefetching and supporting of simultaneous multithreading.

Relevant Graduate Coursework

University of Michigan - Ann Arbor

- EECS 470: Computer Architecture (A+)
- EECS 583: Advanced Compilers (A+)
- EECS 570: Parallel Computer Architecture (A)
- EECS 492: Introduction to Artificial Intelligence (A+)
- EECS 573: Microarchitecture (A)

Awards & Honors

National Scholarship (top 1.8%), *China* 2011
First-Class Scholarship of National IC Talents Training Base, *China* 2012, 2013
First-Class Scholarship for Outstanding Students (top 3%), *Zhejiang University* 2011, 2012, 2013
Honorable Mention in MCM/ICM Contest, *United States* 2013

Skills

Language proficiency: Fluent English, Native Chinese

Programming: Python, C/C++, Bash, LaTeX, Verilog HDL, VHDL, MATLAB

Tools: Caffe2, Torch/PyTorch, TensorFlow, LLVM, Gem5

Teaching Experience

CMOS Integrated Circuits Design Fall 2013
College of Electrical Engineering
Zhejiang University, Hangzhou, China