Saurav Muralidharan

www.sauravm.com ⋄ mail@sauravm.com

RESEARCH INTERESTS

I work on research problems that lie at the intersection of computer systems and machine learning. My current work focuses on improving the performance and efficiency of deep neural networks through pruning & quantization, neural architecture search (NAS), and compiler- and runtime-based approaches.

EXPERIENCE

NVIDIA Research, Santa Clara, CA, USA
- Senior Research Scientist — Oct 2019 – Present
- Research Scientist — Aug 2016 – Sep 2019
- Research Intern — May – Aug: 2013 & 2014

NVIDIA OptiX Team, Salt Lake City, UT, USA
- OptiX Software Intern — May 2011 – Aug 2011

School of Computing, University of Utah, Salt Lake City, UT, USA
- Graduate Research Assistant — Aug 2011 – Jun 2016
- Graduate Teaching Assistant — Jan 2013 – Apr 2013

Indian Institute of Technology – Madras, Chennai, India
- Project Associate — Nov 2009 – Jun 2010

EDUCATION

Ph.D., Computer Science, University of Utah, Salt Lake City, UT, USA — May 2016
- **Advisor:** Prof. Mary Hall
- **Dissertation:** Abstractions and Strategies for Adaptive Programming

B.Tech., Computer Science & Engineering, Kannur University, Kerala, India — Jun 2009

PEER-REVIEWED PUBLICATIONS

Journal Publications

"A Programmable Approach to Neural Network Compression", V. Joseph, **S. Muralidharan**, A. Garg, M. Garland, G. Gopalakrishnan, *IEEE Micro Special Issue on ML for Systems*, 2020.

"Designing a Tunable Nested Data-Parallel Programming System", **S. Muralidharan**, M. Garland, A. Sidelnik, M. Hall, *ACM Transactions on Architecture and Code Optimization (TACO)*, 2016.

"Towards Making Autotuning Mainstream", P. Basu, M. Hall, M. Khan, S. Maindola, **S. Muralidharan**, S. Ramalingam, A. Rivera, M. Shantaram, A. Venkat, *International Journal of High Performance Computing Applications (IJHPCA)*, 2013.

Conference Publications

"Architecture-Adaptive Code Variant Tuning", **S. Muralidharan**, A. Roy, M. Hall, M. Garland, P. Rai, *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2016.

"A Collection-Oriented Programming Model for Performance Portability", **S. Muralidharan**, M. Garland, B. Catanzaro, A. Sidelnik, M. Hall, *ACM Symposium on Principles and Practice of Parallel Programming (PPoPP)*, short paper, 2015.

"Nitro: A Framework for Adaptive Code Variant Tuning", **S. Muralidharan**, M. Shantharam, M. Hall, M. Garland, B. Catanzaro, *IEEE International Parallel & Distributed Processing Symposium (IPDPS)*, 2014.

Workshop Publications

"A Programming System for Model Compression", V. Joseph, **S. Muralidharan**, A. Garg, M. Garland, G. Gopalakrishnan, *NeurIPS Systems for ML Workshop*, 2019.

Theses

"Abstractions and Strategies for Adaptive Programming", Ph.D. Dissertation, University of Utah, 2016.

OPEN-SOURCE SOFTWARE

Condensa: Programmable DNN Compression
`nvlabs.github.io/condensa`

Nitro Automatic Performance Tuning System
`nitro-tuner.github.io`

PATENTS

Bayesian Optimization of Sparsity Ratios in Model Compression
S. Muralidharan, V. Joseph, A. Garg, M. Garland
U.S. Patent Application Number: 62/891,897

POSTERS & TALKS

"Condensa: A Programming System for DNN Model Compression", Talk, *GPU Technology Conference, 2020 (GTC'20)*, March 2020, San Jose, USA

"A Programming System and Automation Libraries for DNN Model Compression", Poster, *Bay Area Machine Learning Symposium (BayLearn 2019)*, October 2019, San Francisco, USA

"Designing a Tunable Nested Data-Parallel Programming System", Invited Conference Talk, *High Performance and Embedded Architecture and Compilation Conference (HiPEAC '17)*, January 2017, Stockholm, Sweden

"Building High-Performance Input-Adaptive GPU Applications with Nitro", Talk, *GPU Technology Conference (GTC '15)*, March 2015, San Jose, USA

"A Collection-Oriented Programming Model for Performance Portability", Poster, *20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*, February 2015, San Jose, USA

"A Framework for Input and Architecture Aware Code Variant Autotuning", Early Research Showcase Poster, *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC '13)*, November 2013, Denver, USA

AWARDS

Winner, University of Utah School of Computing poster competition, 2014
Academic excellence award for ranking 1st in the CS&E department, GCE Kannur, 2009

PROFESSIONAL SERVICE

Conference Program Committee: PLDI 2021, IPDPS (2019, 2018), CC 2019, ICS 2018
Conference External Review Committee: PLDI (2020, 2019), ASPLOS 2019
Conference Review: SC 2016, PPoPP 2016, PLDI 2014, HPCC 2014, ICCS 2013
Journal Review: TACO (2019, 2018), TOPC 2017

TECHNICAL SKILLS

Programming Languages: C, C++, Python
Machine Learning Frameworks: PyTorch, TensorFlow
Parallel Programming Models: CUDA, OpenMP, MPI, Pthreads

GRADUATE COURSEWORK

Computer Architecture, OS, Compilers, Advanced Algorithms, Parallel Programming for GPUs, Advanced Embedded Software, Parallel Computing and HPC, Models of Computation for Massive Data

ACTIVITIES

Graduate Student Advisory Committee, University of Utah, 2012-2014