Saurav Muralidharan

Senior Research Scientist, NVIDIA
✉ mail@sauravm.com ● 🖱 www.sauravm.com ● 📍 San Jose, CA

Overview

Computer scientist with 10+ years of research experience in academia and industry. At NVIDIA Research, my work focuses on two main topics: (1) improving the efficiency, scalability, and correctness of deep neural networks, and (2) using large language models (LLMs) to improve programmer productivity.

Professional Experience

NVIDIA Research Santa Clara, CA, USA
Senior Research Scientist **2016 – Present**

NVIDIA Research Santa Clara, CA, USA
Graduate Research Intern **Summers: 2013 & 2014**

NVIDIA Santa Clara, CA, USA
OptiX Software Intern **Summer 2011**

School of Computing, University of Utah Salt Lake City, UT, USA
Graduate Research Assistant **2011 – 2016**

Indian Institute of Technology (IIT) – Madras Chennai, India
Project Associate **Spring 2010**

Education

Ph.D., Computer Science **2010 – 2016**
University of Utah, Salt Lake City, UT, USA
Advisor: Prof. Mary Hall

B.Tech. (Honors), Computer Science & Engineering **2005 – 2009**
Kannur University, India

Publications

Refereed Conferences & Journals

Uniform Sparsity in Deep Neural Networks, S. Muralidharan, *Sixth Conference on Machine Learning and Systems (MLSys)*, 2023.

A Programmable Approach to Neural Network Compression, V. Joseph, G. Gopalakrishnan, S. Muralidharan, M. Garland, A. Garg, *IEEE Micro Special Issue on Machine Learning 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.

Architecture-Adaptive Code Variant Tuning, S. Muralidharan, A. Roy, M. Hall, M. Garland, P. Rai, *ACM International Conference on Architectural Support for Programming Languages & 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.

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.

Refereed Workshops

Understanding the Effect of the Long Tail on Neural Network Compression, H. Dam, V. Joseph, A. Bhaskara, G. Gopalakrishnan, S. Muralidharan, M. Garland, *Sparsity in Neural Networks Workshop* (*SNN*), 2023.

Efficient Sparsely Activated Transformers, S. Latifi, S. Muralidharan, M. Garland, *ICML Workshop on Dynamic Neural Networks*, 2022 (Spotlight Paper).

Going Beyond Classification Accuracy Metrics in Model Compression, V. Joseph, S.A. Siddiqui, A. Bhaskara, S. Muralidharan, G. Gopalakrishnan, M. Garland, S. Ahmad, A. Dengel, *Sparsity in Neural Networks Workshop* (*SNN*), 2021.

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

Preprints & Other Publications

HighLight: Efficient and Flexible DNN Acceleration with Hierarchical Structured Sparsity, Y. N. Wu, P. Tsai, S. Muralidharan, A. Parashar, V. Sze, J. Emer, arXiv 2305.12718 (2023).

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

Software Systems

Condensa: Programmable Neural Network Compression
`github.com/NVlabs/condensa`

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

Patents

Method to Prune and Accelerate Neural Networks with Hierarchical Fine-grained Structured Sparsity
Y. Wu, P. Tsai, S. Muralidharan, J. Emer
U.S. Patent Application Number: 63/236,629

Bayesian Optimization of Sparsity Ratios in Model Compression
S. Muralidharan, V. Joseph, A. Garg, M. Garland
U.S. Patent Application Number: 16/785,044

Talks & Posters

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

Technical Skills

Programming Languages: C, C++, Python, Swift
Machine Learning Frameworks: PyTorch, JAX
Parallel Programming Models: CUDA, OpenMP

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: NeurIPS (2023, 2022), SC 2016, PPoPP 2016, PLDI 2014, HPCC 2014, ICCS 2013
Journal Review: TACO (2019, 2018), TOPC 2017

Student Mentorship

Cameron Shinn, UC Davis: Ph.D. Intern, Summer 2022
Salar Latifi, University of Michigan: Ph.D. Intern, Fall 2021
Yannan Wu, MIT: Ph.D. Intern, Summer 2021
Vinu Joseph, University of Utah: Ph.D. Intern, Summers 2018, 2019, 2020
Nirmal Prajapati, Colorado State University: Ph.D. Intern, Fall 2017
Thiago S. F. X. Teixeira, UIUC: Ph.D. Intern, Summer 2017

Activities

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