Manas Sahni

sahnimanas.github.io sahnimanas@gmail.com +1-470-309-9496
San Jose, CA

EDUCATION

Georgia Institute of Technology* | Master of Science – Computer Science (GPA: 4.0/4.0)** ***Aug 2019 – May 2021

Specialization in Machine Learning
Research Assistant: Systems for Artificial Intelligence Lab with Prof. Alexey Tumanov
Head Teaching Assistant: Deep Learning, Spring 2021 with Prof. Zsolt Kira, in collaboration with Facebook AI Research

Delhi Technological University* | B.Tech. – Mathematics and Computing (GPA: 3.84/4.0)** ***Aug 2013 – May 2017

EXPERIENCE

NVIDIA, Santa Clara, CA | System Software Engineer, Deep Learning Libraries ***Jul 2021 – Present***

• Libraries to abstract thousands of deep learning primitives optimized for NVIDIA GPUs with optimal performance & size.

NVIDIA, Santa Clara, CA | Software Engineering Intern, TensorRT ***May 2020 – Aug 2020***

• Performance optimizations of deep-learning recommender systems by profiling, identifying bottlenecks, and designing & implementing system modules across NVIDIA's GPU product line
• Contributed CUDA and multi-threading solutions to enable 2x improvements in performance and multi-GPU scalability into NVIDIA's official winning entry to the cross-industry MLPerf Inference benchmark.

Samsung R&D, India | Machine Learning Software Engineer ***Aug 2017 – Jun 2019***

• R&D aimed at enabling deep-learning efficiency for applications in mobile & low-power systems via neural architecture design, low-precision quantization, pruning/compression, distribution on heterogeneous hardware, code optimization. Research output published via patents and papers.
• Directly helped enable over 15 USP camera features deployed on flagship Galaxy S9 & S10 phones. Contributed upto 20x optimizations for speed, memory, and battery shipped in the Samsung Neural SDK.

Samsung R&D, India | Computer Vision Intern ***Jun 2016 – Jul 2016***

• Interned with CTO group's Advanced Technologies Lab. Studied hand-crafted image features & scoring measures to automate process of video highlighting & summarization. Implemented algorithm in C++ using OpenCV and Eigen-C++

RESEARCH PROJECTS

CompOFA – Fast Neural Architecture Search for Diverse Hardware
Insights to enable a 200x faster and 2x cheaper technique for neural architecture search for efficient deployment on diverse hardware. First-author of conference paper published at ICLR 2021.

Dynamic Multi-Stage Model Cascades for time-sensitive clinical workflows
Framework for optimization of time/cost-constrained sequential ML pipelines. Applied towards septic shock prediction in ICU patients to achieve a 19.6x cost reduction and 26.1 hours earlier prediction without loss of accuracy.

Soft Real-Time Machine Learning (SRTML)
An open-source research framework for declaratively-specified machine learning inference pipelines to automate model selection, hardware selection, and configuration for end-to-end performance across all participants in ML ecosystem.

PATENTS & PUBLICATIONS

• M.Sahni, S. Varshini, A. Khare, A. Tumanov, *"Compound Once-For-All Networks for Faster Multi-Platform Deployment"*, **International Conference on Learning Representations (ICLR) 2021**
• M. Sahni, A. Abraham, S. Allur, V. Mala, *"Method and electronic device for handling a neural model compiler"*, **US Pending Patent** US20200065671A1, filed 23 August 2018
• A. Abraham, M. Sahni, and A. Parashar, *"Efficient Memory Pool Allocation Algorithm for CNN Inference"*, *IEEE* **International Conference on High Performance Computing (HiPC), 2019**
• *Workshop Poster:* B. Singh, M. Sahni, and S. Allur, *"Shunting Connections in MobileNet v2"*, **NeurIPS Workshop on Machine Learning on the Phone and other Consumer Devices (MLPCD 2), 2018**

AWARDS & ACTIVITIES

• Blog on efficient deep learning, ***EfficieNN*, with reach of over 60k** and featured by *HackerNews & DL Weekly Newsletter*
• **Samsung Young Achiever of the Year**, 2018-19***;* Samsung Citizen Award for Technological Excellence,** presented for performance optimization of 3D face-reconstruction algorithms used on Galaxy S9 & Note9 smartphones
• Pesented talk titled ***"Challenges in Embedded ML and influence on vision solutions"***, at Indian Institute of Technology (IIT) Guwahati
• Volunteered training and project mentoring in machine-learning for community college students; volunteered training in public-speaking for high-school students in India.

TECHNICAL SKILLS

• **Programming & Scripting:**	Proficient in C++, Python, MATLAB, Android NDK, SQL, Git, Shell, Docker
• **Deep Learning:**	Convolutional Neural Nets, Neural Architecture Search, Transformers, PyTorch, TensorFlow, Keras
• **Systems & Performance:**	CUDA, MPI, OpenMP, OpenBLAS, Boost-C++, Halide, OpenCL