Manas Sahni

https://sahnimanas.github.io sahnimanas@gatech.edu +1 (470)-309-9496
Atlanta, GA

EDUCATION

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

Research Assistant: Systems for Artificial Intelligence Lab with Prof. Alexey Tumanov
Teaching Assistant: Deep Learning, Spring 2019 with Prof. Zsolt Kira

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

EXPERIENCE

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

- Drove 2x performance improvements and multi-GPU scalability for NVIDIA's official submission to the cross-industry MLPerf Inference benchmark, in the recommender system category.
- Profiled and identified bottlenecks, followed by design and implementation of key optimizations on DLRM recommender inference, through system modules powered by CUDA & multi-threading.

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

- *Samsung Young Achiever of the Year (2018-19); Samsung Citizen Awardee for Technology Excellence (2018)*
- R&D aimed at enabling deep-learning applications on low-power smartphones/embedded systems via neural architecture design, low-precision quantization, pruning, distributing on heterogeneous hardware, and system optimization.
- Contributed upto 20x optimizations for speed, memory, and battery shipped in the Samsung Neural SDK. Directly helped meet performance targets on over 15 USP camera features deployed on Galaxy S9 & S10 phones.

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

- Partnered with CTO group's Advanced Technologies Lab. Studied hand-crafted image features & scoring measures to generate video summaries. Implemented algorithm in C++ using OpenCV and Eigen

SELECTED PROJECTS

CompOFA – Fast Training of Neural Networks for Diverse Hardware

- [Ongoing] Improving a state-of-the-art neural architecture search technique for deployment on diverse latency requirements in one-shot training, with an emphasis on reducing training time & carbon emissions by 2x. Advised by Prof. Alexey Tumanov.

Soft Real-Time Machine Learning (SRTML)

- Helping develop an open-source research framework for declaratively-specified machine learning inference pipelines with latency constraints and automate their model selection, hardware selection, and configuration for end-to-end performance.

Anatomy of a High-Speed Convolution

- Developed a tutorial on how production-level deep learning libraries employ concepts from high-performance and parallel computing, replicating OpenBLAS performance of 100x speedup on GEMM.

PATENTS & PUBLICATIONS

- *Patent:* 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
- *Conference Paper:* 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
- 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
- **Machine Learning:** Convolutional Neural Nets, RNNs, Caffe, PyTorch, TensorFlow, ONNX, Android NN-API
- **Systems & Performance:** CUDA, OpenBLAS, Boost-C++, Halide, OpenCL