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** ***Aug 2019 – May 2021***
Key Coursework: Deep Learning, Artificial Intelligence, Data & Visual Analytics

Delhi Technological University **| B.Tech. – Mathematics and Computing (GPA: 3.84/4)** ***Aug 2013 – May 2017***
Key Coursework: Operating Systems, Computer Architecture, Numerical Linear Algebra, Applied Graph Theory

EXPERIENCE

Samsung Research | 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 at the intersection of ML & systems, aimed at improving efficiency of deep-learning applications on low-power smartphones/embedded systems.
- Contributed upto 20x optimizations for speed/memory/battery on over 15 USP camera features. Directly helped meet performance targets for deployment on Galaxy S9 & S10 phones.
- Represented org at Qualcomm, San Diego for S/W integration of dedicated ML hardware; led early efforts for critical accuracy fixes and developer API design.
- Key skills practiced: convolutional neural networks, parallel processing, model compression, ML deployment, edge computing

Samsung Research | Summer Intern ***Jun 2016 – Jul 2016***

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

Ernst & Young LLP| Summer Intern ***Jun 2015 – Jul 2015***

- Data analysis intern; assisted TV broadcaster clients in identifying potentially fraudulent franchisees using revenue data & fuzzy string-matching. Applied anomaly detection methods to find evidence of collusion and devise correction strategies.

SELECTED PROJECTS

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.

Offline Neural Network Compiler

- Devised a novel method to profile and optimally allocate neural network models in an embedded heterogeneous setting. Outcome ***realized as a patent application*** pending with US and India Patent Offices.

Deep Reinforcement Learning & Evolution Strategies for Game-Playing

- Studied the use of evolutionary strategies as scalable alternatives to deep Q-learning for game-playing on Atari from raw pixels

Multi-task CNNs for Face Analysis

- Implemented & extended the HyperFace Multi-Task CNN for unified prediction of face presence, landmarks, gender & identity.

PATENTS & PUBLICATIONS

- *Patent:* M. Sahni, A. Abraham, S. Allur, V. Mala, **"Method and electronic device for handling a neural model compiler"**, US & India Pending Patent (2018141031660), filed 23 August 2018
- *Conference Paper:* A. Abraham, M. Sahni, and A. Parashar, **"Efficient Memory Pool Allocation Algorithm for CNN Inference"**, (forthcoming) *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 40k+** 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, TensorFlow, PyTorch, ONNX, Android NN-API
• **Systems & Performance:**	Code optimization, heterogeneous/parallel systems, OpenBLAS, Boost-C++, Halide, OpenCL