Gautam Kumar

(+91)9430425857 | gautam1994kumar@gmail.com | gautamgtm | gautamgtm | gautamgtm

Education

Indian Institute of Technology Roorkee, India *GPA: 8.46 / 10*

B.Tech. in Electrical Engineering *July'11 - May'15*

Experience

Qualcomm *San Diego, USA*

Machine Learning Engineer *Sep'22 – Present*

- Designed, developed, debugged, and tested the Qualcomm Neural Processing (QNN) SDK, enabling efficient machine learning on edge devices.
- Architected and implemented profiling APIs for QNN, including support for opaque objects, continuous profiling, and multi-graph profiling.
- Designed and developed APIs to support multi-branch GenAI LLM models.
- Led the profiling and signal efforts for QNN across Qualcomm.
- Owned and maintained API compliance tests for QNN, ensuring consistent and reliable functionality for all developers.

Qualcomm *Hyderabad, India*

Machine Learning Engineer *March'18 – Sept-22*

- Developed, debugged, and tested Snapdragon Neural Processing Engine (SNPE) SDK, enabling machine learning on edge devices.
- Optimized hardware-accelerated inferences for power and performance across CPU, GPU, and DSP runtimes on Qualcomm chipsets.
- Worked extensively on deep learning model conversion of Tensorflow, Caffe, Onnx, and PyTorch pre-trained models to Qualcomm Proprietary format for SNPE SDK.
- Developed and maintained the performance infrastructure of SNPE SDK.
- Analyzed and debugged the power and performance of neural networks. Led the development of power and performance testing framework for SNPE SDK, ensuring high reliability and consistency in performance evaluation.

Qualcomm *Hyderabad, India*

Chipset Power Systems Engineer *June'15 – Feb'18*

- Led the roadmap for power management features (hardware and software) of Snapdragon processors, coordinating cross-functional teams to deliver innovative solutions.
- Chipset Power Lead of Snapdragon 425 SoC and Snapdragon 632 SoC.
- Designed power grids to achieve optimal power efficiency within cost constraints. Tuned clock plans for multiple cores and sub-systems.
- Modeled and projected power usage (average and peak) goals for Snapdragon processors, aligning with product performance and efficiency targets.
- Investigated hardware and software architectures, analyzing data flow for key use cases to identify novel power optimization solutions.
- Enhanced CPU DCVS and scheduling mechanisms, achieving significant power and performance enhancements.
- Explored and developed machine learning-based solutions for improving the power and performance of Snapdragon chipsets.

Qualcomm *Hyderabad, India*

Software Engineering Intern *May'14 – July'14*

- Developed a fully automated Testing Framework (JTF) for Qualcomm-powered Android devices, streamlining the testing process for efficiency and simplicity.
- Automated NFC Testing using a robotic arm and integrated it into the JTF.

Projects

Evaluation of Transformer Neural Network Models *Qualcomm*

Deep Learning *2019*

- Evaluated Transformer based models for sequence to sequence learning and comparison with the existing state of the art RNN Models.
- Scoped out the Tensorflow Layers Implementation to support the model in SNPE.

Active Noise Cancellation using Recurrent Neural Networks *Qualcomm*

Deep Learning *2018*

- Implementation of a Recurrent Neural Network for audio noise suppression based on RnnNoise project by Mozilla.
- Developed an Android application which uses Tensorflow Mobile for execution of the neural networks.

Power Waveform Analyzer for Energy Residency analysis *Qualcomm*

Power Systems *2018*

- Development of a tool for analysis of power waveforms collected on mobile chipsets.
- Improved the methodology of use-case power analysis based on the energy consumption patterns across the available frequency levels for CPU, GPU, Memory, Modem, and, other cores.

Temperature Prediction using Recurrent Neural Network — *Qualcomm*

Deep Learning — *2017*

- Developed a Hierarchical Bi-directional Recurrent Neural Network Architecture to predict the future temperatures of the most thermally-sensitive cores inside a Snapdragon Processor, for pro-active thermal mitigation to allow the device to run near the thermal limit.
- Inputs to the network comprised of a sequence of SoC states over the past few seconds, where each state is defined by 44 features consisting of low-level CPU parameters and readings from temperature sensors.
- Improved the robustness of the model by Ensemble Averaging and K-Fold Cross Validation.

Facial Keypoints Detection using Convolutional Neural Network — *Kaggle*

Deep Learning — *2017*

- Implemented a Deep Convolutional Neural Network for prediction of keypoints positions on the human face which could be used for various applications such as Facial Recognition.
- Input consists of thousands of B/W images of 96x96 pixels.

Rainfall Estimation using Recurrent Neural Network — *Kaggle*

Deep Learning — *2016*

- Implemented a Recurrent Neural Network for prediction of Hourly Rainfall gauge levels recorded over a few months in 2014 over the US mid-western corn-growing state.
- Inputs consisted of a sequence of multiple polarimetric weather radar observations over the course of an hour, where each measurement consisted of 22 features.
- Employed Ensemble Averaging and K-Fold Cross Validation to improve the prediction accuracy.

Workload Classification using Supervised Learning — *Qualcomm*

Machine Learning — *2016*

- Developed a novel approach for classification of CPU workloads into two disparate classes for improving CPU Governor Algorithm.
- Used k-Nearest Neighbours Algorithm for classifying new workloads.

Lowlevel CPU stats Logger for ARM CPU — *IIT Roorkee*

CPU Architecture — *2016*

- Developed a logging mechanism for periodically collecting lowlevel CPU stats like Instructions Executed, Cache Accesses, DDR Accesses and Activity in real time on Snapdragon Chipsets having ARM-based Apps Processor.

Implementation of Backpropagation Algorithm in VHDL — *IIT Roorkee*

Hardware Programming | Machine Learning — *2015*

- The project envisages the performance of FPGA (Field Programmable Gate Array) for applications in machine learning by implementing Back-propagation Algorithm and compare the execution time with the software implementation in python.
- Implemented the Backpropagation algorithm in VHDL. Constructed different modules for the hidden nodes and output nodes which give the flexibility to constructing any network.
- Verified the accuracy and correctness of our implementation by simulating the XOR problem as a small dataset problem and Fisher Iris problem as a large dataset problem.

Letter Image Recognition using Neural Network — *Qualcomm*

Machine Learning — *2014*

- The project involved the development of an artificially intelligent method to recognize the hand-written English alphabets.
- In the development process, 20000+ Character images were used, based on 20+ different fonts and each letter was randomly distorted to produce a file of 20,000+ unique stimuli, each having 15+ primitive numerical attributes which were used to train the neural network.

Coursework

Undergraduate — *IIT Roorkee*

Machine Learning | Artificial Neural Networks | Computer Systems & Programming | Data Structures | Advanced System Engineering | Control System

Independent

Deep Learning | Operating Systems | Computer Architecture | Algorithms | Introduction to Computer Science and Programming

Skills

Programming & Machine Learning

C++ | C | Python | MATLAB | Tensorflow | Caffe | Keras | Onnx | Git | GDB

Platforms and Build System

Linux | OpenEmbedded Linux | Android | CMake | Ninja | Docker

Scholarships & Achievements

2019	**Patent: US 10331195**, Power and Performance Aware Memory-Controller Voting Mechanism.	*Qualcomm*
2019	**IEEE: 2018 19th ISQED**, Power and Performance Aware Memory-Controller Voting Mechanism.	*Qualcomm*
2014-16	**Recipient**, Multiple Qualstar Hall of Fame for exceptional contributions as an Engineer and Intern.	*Qualcomm*
2011-15	**Recipient**, Merit-Cum-Means Scholarship for undergraduate studies (INR 25000/year)	*IIT Roorkee*