GOKULKRISHNA MUTHUSAMY

New York City | +91 9629925795 | gokulkrishna98@gmail.com

Incoming graduate student at NYU MS CS with over 2 years of work experience. Looking for an internship starting May 2024.

[LinkedIn] [Personal Website]

Education

New York University (Courant), New York, NY — Sept 2023 - May 2025
Master of Science in Computer Science

National Institute of Technology (NIT), Tiruchirappalli, TN — Jun 2020
Bachelor of Technology in Computer Science and Technology — 8.6/10

Experience

Senior Engineer, *Samsung Research India*, Bangalore, KA — Apr 2022 – Aug 2023
SNAP - CPU, GPU and Compiler team for Neural Acceleration.

- Programmed kernels for complex neural layers like group normalization and updated MLIR to support TfLite model conversion, enabling complete execution of Stable Diffusion models on Mobile GPUs, resulting in 230% improvement in performance over CPU.
- Designed an automatic recompilation & caching tool for OpenCL kernels on GPUs to reduce caching overhead and prevent crashes during GPU driver updates, which resulted in a 25% reduction in crash reports and a 20% increase in the productivity of engineers and the use case team. Received Spot award in Q3 2022 for developing this tool.
- Accelerated 10 use cases on Samsung flagships (Galaxy S23) to improve user experience, achieved a 20% improvement in load time and 10% in execution time and overall performance improvement of 1.4x (S23) and 1.3x (Fold5).

ML Software Engineer, *Samsung Research India,* Bangalore, KA — Jan-2021 – Apr 2022
SNAP - CPU, GPU and Compiler team for Neural Acceleration.

- Engineered a Profiling tool for calculating the layer-wise performance of kernels in ArmNN during inference of ML model. Resulting in the faster diagnosis of performance degradation and improved the productivity of the team by 30%.
- Collaborated with Galaxy RAW (USP) team to solve greenish tinge issues on images from AINR (noise reduction) by implementing new normalization method to support float16 quantized precision execution. This improved inference time by 34% over the previous method.
- Directly helped in enabling over 15 USP camera and gallery features deployed on Samsung Galaxy S22. Contributing to over 1.5x optimizations for speed, memory and battery.

Intern, ***Samsung Research India,*** Bangalore, KA — May 2019 – July 2019
Keyboard Intelligence - OnDevice AI

- Improved the emoji prediction in the Samsung keyboard by incorporating keystroke statistics in the NLP sentiment analysis algorithm, resulting in a 5% improvement in accuracy on personalized datasets.

Patents & Publications

A1 graded patent by Samsung Research HQ — Oct 11, 2022
The main inventor of Systems and Methods for On-Device Validation of a Neural Network Model, this idea reduces validation memory requirement by 100% and makes the computations on-device friendly. Inventors- M Gokulkrishna, Siva Kailash, R. Prasanna, Rajath Elias, Ashok Kumar, Praveen. Pending patent: IN202241058188

Technical Skills

Programming Languages: C++, Python, OpenCL, CUDA (platform), C, JavaScript, Shell, Rust.
Technologies: TensorFlow, TfLite, ArmNN, Pytorch, ONNX, Apache TVM, MLIR, Quantization, Transformers.
Tools: Git, Perfetto, Gdb, Docker, FlatBuffers, Bazel, Valgrind, Asan, Hwasan, Android NDK.

Contributions & Activities

- Active contributor to TensorFlow open-source community.
- Lead the design team for Vortex, Computer Science department in NIT Tiruchirappalli.
- Received Full Stack Web Developer Nanodegree from Udacity [credentials]