

Professional Experience

AI Software Solutions Engineer **Intel Corporation** **January'24 – Ongoing**

- Developing **high-performance kernels** for deep learning operators on the upcoming Intel GPU using SYCL. Conducting runtime analysis and optimization to enhance latency, memory bandwidth, I/O access, and compute utilization, and adding **symbolic support** to handle **dynamic shapes** of the operators.
- Enhancing the software stack by creating an optimized graph in C++ to handle **complex operations** and utilizing **MLIR** types, attributes, and dialect operations.

Senior ML Engineer **Qualcomm Corporate R&D** **December'22 – January'24**

- Spearheaded **ONNX optimizations** for NLP and CV models on Qualcomm's **AI100 accelerator**, achieving a notable **8.5% performance boost** for large language models (LLMs) like **ChatGLM2-6B** through C++ **node fusion** of layer-normalization modules.
- Doubled the efficiency** of NLP transformer decoder models (**OPT-LLM, GPT variants**) by implementing key optimizations, including **caching Key-Value matrices** and minimizing DDR reads and writes.
- Developed a Graph Neural Network algorithm to enhance compiler efficiency, resulting in a filed patent.
- Led a three-member team** in optimizing and deploying the top 120 models from Hugging Face library.

ML Engineer **Qualcomm Corporate R&D** **November'20 – November'22**

- Engineered software modules in C++ & Python for AI/Deep Neural Network frameworks.
- Implemented auto-detection of post-processing components for image classification and object detection models, replacing them with **optimized kernels** to improve model accuracy during **quantization**.
- Achieved a **28.2% performance improvement** for NLP encoder models (**BERT and variants**).
- Enhanced operator support in the **GLOW compiler** for the **Cloud AI100 SDK**.

Research Assistant **IIT-Bombay** **August'20 – October'20**

- Developed transformer-based architecture exploring the relationship between video, audio, and text features.

Patent and Publication

U.S. Patent application 18/330,253 and 18/500,014 (Pending)

- Proposed "Pre-Processing For Deep Neural Network Compilation Using Graph Neural Networks," filed on June 06, 2023.

Home Automation Using Panoramic Image Using IoT  (Published in 2018)

- International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE).

Tech Stack for Software Development and Machine Learning

- Programming:** Python, C++, SYCL (DPC++)
- Machine Learning Frameworks:** PyTorch, ONNX, ONNX Runtime
- ML Domain & Techniques:** NLP, CV, Quantization, Pruning, Node Fusion, Graph Optimization, GNN
- Others:** GPU Optimization, Git, Docker, GLOW (Machine Learning Compiler)

Education

Mumbai, IN **IIT-Bombay ** **July'18 - July'20**

- M.Tech in Computer Science and Engineering, July 2020. CPI: **8.43** (on scale of 10).

Mumbai, IN **University of Mumbai ** **June'14 – June'18**

- B.E. in Computer Engineering, June 2018. CPI: **9.07** (on scale of 10).

Master Thesis

- Computational Model to Understand and Predict Emotions. (2020)