

Professional Experience

Senior ML Engineer	Qualcomm Corporate R&D	November'20 – Ongoing
---------------------------	-----------------------------------	------------------------------

- Worked on ONNX optimizations for NLP (Natural Language Processing) and CV (Computer Vision) models for faster inference on Qualcomm's AI100 accelerator.
 - Implemented node fusion of layer-normalization module into a single kernel in C++ for large language models (LLMs - ChatGLM2-6B), resulting in 8.5% boost in the performance (number of inferences/second).
 - Improved performance of NLP transformer decoder models (OPT - LLM by Meta, and other GPT variants) by 2x by caching the Key-Value matrices of the attention layer and minimizing DDR reads & writes.
 - Improved performance of NLP encoder models (BERT and its variants) by 28.2% by node fusion of attention module and Graphcore's packing strategy (specifically designed for QnA tasks) .
- Designed and implemented software modules for Artificial Intelligence/Deep Neural Network frameworks and tools in C++ & Python automating general ONNX graph optimizations.
 - Implemented auto-detection of post processing part for Image classification, and object detection models, and replaced it with optimized kernels to improve the accuracy of the model during quantization.
 - Implemented Graph algorithms for sorting nodes and removing unused nodes in a graph for faster inference.
- Developed Graph Neural Network (GNN) based algorithm to improve the compiler efficiency and filed patent.

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

- Worked in joint collaboration of IBM and IIT Bombay on Understanding emotions in Sarcasm.
- Trained a transformer based architecture for leveraging the relation between video, audio and textual features.
- Proved that emotion information was necessary to identify sarcasm more precisely. Experiments with emotion information had 15.6% better performance.

Patent and Publication

U.S. Patent application 18/330,253 (*Pending*)

- "Pre-Processing For Deep Neural Network Compilation Using Graph Neural Networks", 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++
- Machine Learning Frameworks: PyTorch, ONNX, ONNX Runtime.
- ML Domain & Techniques: NLP, Computer Vision, Quantization, Pruning, Node Fusion, Graph Optimization.
- Others: Git, Docker, GLOW (Machine Learning Compiler), AWS.

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)
 - Created dataset 'emo-UStARD' by annotating 'MUSARD' with 8 primary emotions, arousal & valence.
 - Conducted experiments exploring every aspect of textual modality & observed 18% increase in accuracy score in multi-label Emotion Prediction when additional information is passed.