in
linkedin.com/in/prasanna-biswas

Prasanna Biswas

(+91) 9922365239
prasanna.biswas14@gmail.com

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.
  - Improved performance of NLP encoder models by 28.2% by node fusion of attention module.
  - For Large Language Models (LLMs) like ChatGLM2, used node fusion for layer-normalization into a single kernel (in C++) supporting both PyTorch and AI100 backend, resulting in 8.5% boost in the performance.
  - Improved performance of NLP tranformer decoder models by 2x by caching the Key-Value matrices of the attention layer and using mixed-precision (FPl6 along with model quantization).
- Designed and implemented software modules or components for Artfical 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 model with self-attention and cross-attention of video, audio and textual features.
- Result - 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 (High Performance Runtime for ONNX Models supporting most machine learning frameworks and hardware backends).
- 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 Emotions in Sarcasm. (2020)
  - Created dataset 'emo-UStARD' by annotating 'MUStARD' 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.