



Prasanna Biswas

AI Software Solutions Engineer at Intel Corporation



Work experience

present
↑
Jan 2024

AI Software Solutions Engineer

Kernels, Falcon Shores, Intel Corporation

- Developing **high-performance kernels** with dynamic shape support for Intel's next-gen GPU using SYCL, optimizing latency, memory bandwidth, I/O access & compute utilization.
- Programmed an efficient **cumsum kernel**, achieving **2x perf improvement** over IPEX eager mode implementation.
- Designed and **implemented** complex operations like **TopK** and media operators such as **Brightness and Contrast** as graphs in C++ using MLIR types and attributes, enabling efficient GPU execution.
- Innovated a novel machine learning algorithm combining **VAEs** and **Diffusion Models** for NLP and CV.
- **Co-authored two papers**; one **submitted to CVR 2025** and second **submitted to IEEE CONNECT-2025**.

Jan 2024
↑
Dec 2022

Senior ML Engineer

ML Applications, Cloud AI100, Qualcomm CR&D

- Spearheaded **ONNX optimizations** on Qualcomm's **AI100 accelerator**, achieving an **8.5% performance boost** for large language models (LLMs) like **ChatGLM2-6B** through **node-fusions, graph simplifications**.
- Enhanced **GPT model** efficiency by **2x** through caching Key-Value matrices and minimizing DDR reads/writes.
- Designed 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.

Nov 2022
↑
Nov 2020

ML Engineer

ML Applications, Cloud AI100, Qualcomm CR&D

- Engineered software modules in C++ & Python.
- Introduced auto-detection of post-processing in CV models, replacing them with **ABP & NMS** optimized kernels for **80% improvement** in quantization accuracy.
- Achieved a **28.2% perf improvement** for (**BERT** and variants) through **Graphcore's packing strategy**.
- Enhanced operator support in the **GLOW compiler** for the Cloud AI100 SDK.



Patent and Publications

Mar 2025 • **Efficient Deep Learning Model Architecture for Emergence of Machine Style Calligraphy**

IEEE CONNECT-2025 Conference (Submitted)

Dec 2024 • **Machine-Style Handwriting Generation with Diffusion based latent generation**

CVR 2025 Conference (Accepted for presentation)

Jun 2023 • **Pre-Processing For Deep Neural Network Compilation Using Graph Neural Networks**

USPTO: 18/330,253 and 18/500,014 (Pending)

Jun 2018 • **Home Automation Using Panoramic Image Using IoT**

Published in: 2018 ICRIEECE



Contact



Email

prasanna.biswas14@gmail.com



Phone

(+91) 9922365239



Profile

/in/prasanna-biswas



Portfolio

prasannabiswas-iitb.github.io



M Tech, Thesis

Computational Model to Understand Emotions in Sarcasm

Created the 'emo-UStARD' dataset by annotating 'MUSARD' with 8 emotions, arousal, and valence.

Conducted experiments, observing an **18% increase** in accuracy across various aspects of textual modality.



Technical Blogging & Content Creation



Technical Blogs

GPUs and CUDA Programming



YouTube Channel Co-Owner & Python Instructor

Successfully manage a channel with 1.5k+ subscribers.



Technologies

Programming:

- Python, C++
- GPU: SYCL(DPC++), CUDA

Machine Learning Frameworks:

- PyTorch
- ONNX, ONNX Runtime

ML Domain & Techniques:

- NLP, CV
- Graph Optimization, GNN
- Quantization, Pruning, Node Fusion

Others:

- GPU Optimization
- Git, Docker
- GLOW (Machine Learning Compiler)



Education

M Tech, 2020

- IIT Bombay
- CPI: 8.43/10

B Tech, 2018

- VESIT, Mumbai
- CPI: 9.07/10