Aakash Varma Nadimpalli

Portfolio: aakashvarma.com

in LinkedIn
@varmology

Education

- **Birla Institute of Technology and Science, Pilani (WILP)** — India
  Master of Technology - Data Science and Engineering; GPA: 9.32 — *2022 - 2024*
- **Vellore Institute of Technology** — India
  Bachelor of Technology - Electrical and Electronics Engineering — *2015 - 2019*

Skills Summary

- **Languages**: Python, C, C++
- **Frameworks**: PyTorch, ONNX, TensorFlow, Caffe, MLIR, TVM
- **Techniques**: Quantization (AWQ, GPTQ, GGUF/GGML, SmoothQuant), LoRA, QAT, PTQ, GEMM
- **Models**: LLMs (DeepSeek, Qwen, LLama, CLIP, etc), Diffusion Models (SDXL, CogVideoX), Segmentation, Detection and Classification Models (ResNets, YOLOs, RCNNs, etc)
- **Mathematics**: Linear Algebra, Calculus, Approximation Algorithms, Applied Numerical Methods, Cost Function Modeling

Experience

- **Dheyo AI** — Remote
  Staff Software Engineer — *Aug 2024 - Present*

- **Oxmiq Labs** — Hyderabad, India
  Staff Software Engineer — *May 2024 - Jul 2024*
  - **Compiler & Runtime Development**:
    - Architected a PyTorch compiler that leverages torch.compile to lower operations to TOSA instruction set via torch-mlir
    - Engineered an asynchronous queuing system for torch dispatch that enables efficient execution on Tenstorrent Hardware
    - Developed an optimized garbage collection system for on-device tensor deallocation in torch's eager execution mode during runtime
    - Implemented a comprehensive self-validating framework with 3-point validation between TOSA and torch operations
  - **Performance Optimization**:
    - Designed algorithms for efficient matrix multiplication tiling optimized for custom hardware acceleration
    - Researched CUDA/CuBLAS tiling strategies to enhance tensor operation performance
    - Analyzed batching effects in MLP and Attention layers of transformer architectures to optimize inference
  - **Model Development**:
    - Evaluated Deepseek_r1_Distil_Qwen_1.5_b and CogVideoX models, creating optimized inference pipelines
    - Implemented Llama 3.2 1B from scratch with custom optimizations for deployment

---

- **Kinara, Inc.** — Hyderabad, Telangana, India
  Various Roles — *2020 - 2024*
  Staff Engineer — *2023 - Apr 2024*
  - **LLM Deployment and Optimization**:
    - Deployed Large Language Models on Kinara's Edge AI Processor through novel quantization techniques and compiler optimization
    - Examined outliers in LLM architectures like llama7b, qwen7b, and tinyllama for ARA-2 deployment
    - Analyzed quantization methods (AWQ, GPTQ, GGUF/GGML) for optimizing LLMs
    - Developed a framework for LLM Smoothing using a modified version of QmniQuant that incorporates smoothing into the down projection layer of attention blocks
    - Implemented FlashAttention tiling and SoftMax online normalization calculator for memory-efficient precise attention mechanism on ARA-2 NNP
    - Analyzed & pruned LLM layers using SVD and Block Importance, enhancing throughput from 2 to 9 Tokens/sec (4.5x improvement) with minimal accuracy loss on lm-eval
    - Employed LoRA, QLoRA, & LoRA+ techniques to restore pruned models to SOTA accuracy on lm-eval
    - Developed a Knowledge Distillation framework using FSDP across multiple GPUs (A10, H100, A100)
  - **CLIP Model Enhancement**:
    - Optimized OpenAI's CLIP model by enhancing its Transformer blocks, focusing on KQV projection layers through quantization observer analysis
    - Investigated the impact of quantization errors in mean, variance, and inverse square root in Layer normalization within Transformer blocks
    - Analyzed systematic outliers in hidden layer features, developing new quantization computation to mitigate errors

Senior Software Engineer *2021 - 2023*

- **Quantization Research**:
  - Conducted comparative analysis of rounding techniques (Ada-Round, RNE, RAI) for ResNet50: Original PyTorch model achieved 76.84% accuracy, RAI platform yielded 74.00%, RNE platform attained 76.16%, and Ada-round Simulator matched PyTorch's accuracy at 76.84%
  - Invented a novel Inverse Square Root Approximation for neural network normalization layers with 90% reduction in MSE and 83% reduction in MAE compared to existing techniques
  - Analyzed impact of observer types on rounding techniques during Quantization Aware Training (QAT)
- **Hardware Optimization**:
  - Implemented distributed online normalizer method for efficient bilinear interpolation on ARA-1 NNPs
  - Developed efficient tiling method for optimized tensor permutations on ARA-2 hardware
  - Enhanced YOLOv5 performance through activation distribution analysis and systematic offsetting, improving quantized model precision from 47.8% to 51.6%
- **Framework Development**:
  - Refactored the compiler framework to support multiple target hardware platforms
  - Contributed to the development of the ARA-2 simulator
  - Evaluated Qualcomm's AIMET and Intel's NNCF frameworks for in-house quantization framework development
  - Built conversion framework for PyTorch QAT models (JIT) to ONNX QDQ format (Graph model)

Software Engineer *Jun 2020 - 2021*

- **Kernel Development**:
  - Developed optimized kernels for powers-of-two approximation exponentiation in SoftMax functions
  - Created approximation functions for neural network operations (Swish, Mish, GeLU) using piecewise techniques
  - Engineered exp operation approximation for ASIC using Applied Numerical Methods
  - Designed efficient kernels for neural network operations including convolutions, deconvolutions, innerproducts and layernorms for various network architectures
- **System Development**:
  - Developed an AI compiler based on the Caffe framework
  - Created an ensemble machine learning algorithm to predict ARA-1 chip power consumption
  - Formulated mathematical cost functions to predict kernel cycle costs aligned with ASIC performance metrics
  - Designed precision-preserving mathematical kernels for complex operations like ROIAlign and bilinear interpolation on ASIC

- **Wipro** Bengaluru, Karnataka, India
  ML and Big Data Developer *Jul 2019 - Jun 2020*
  - **Data Processing**:
    - Implemented big data processing pipelines for large-scale machine learning applications

Research

- **MTech Thesis**: Fine-Tuning and Quantization Techniques for Enhanced Efficiency in LLMs for Task-Specific Code Generation
- **BTech Research**: Convolutional Neural Networks based Dementia and Tumor Classification from MRI Brain Images (Published in IEEE Xplorer)
- **The RoPE Compatibility Problem in Multi Head Latent Attention**: A Detailed Mathematical Overview
- **Effects of Batching in Transformer Blocks**: Analyzed MLP and Attention blocks on the effects of batching
- **Inverse SQRT Approximation using Range reduction and Piece wise Approximation**: Developed efficient approximation technique achieving 90% reduction in MSE and 83% reduction in MAE

Honors and Awards

- Winners of Startup Street 4.0 — VIT Vellore (Jan 2018)
- Wolfram Award at DevSpace — Wolfram Alpha (Jan 2018)
- Best ML Implementation at Hackoverflow — CSED (2018)

Leadership Experience

- **CSED/SHEILD at VIT** VIT University
  Founding Member & Vice President of Tech and Design *2016 - 2019*
- **Venturesity VIT** VIT University
  Vice President of Design *2017 - 2019*