V S Navneet Kanna

navneetkanna04@gmail.com
Bangalore, India
https://www.linkedin.com/in/navneet-kanna-099487221/
https://github.com/NavneetKanna
https://navneetkanna.com

Research Interests

GPU kernels, LLM inference systems, hardware-aware ML systems, ML Compilers

Professional Experience

Aug 2023 – present
Bangalore, India

Technical Lead - Machine Learning Engineer
XenReality Technologies Private Limited

- Built and deployed a production GPU inference infrastructure on NVIDIA GPUs, serving **over 30,000 requests per day** through a dedicated inference server and backend request pipeline.
- Led the end-to-end R&D, development, and deployment of a production vision AI system, securing a **seven-figure (INR) commercial contract**.
- Optimized and deployed deep learning models to edge devices, enabling low-latency, real-time inference for production vision AI applications.
- Collaborated directly with customers through on-site visits to understand workflows, identify pain points, gather requirements, and deliver production AI solutions.

Education

Sep 2022 – Aug 2023
Glasgow, Scotland

Master of Science, Robotics and AI
University of Glasgow
Dissertation: A5 (Excellent) - "Modelling of Superconducting Flux Pumps using AI Techniques"

Sep 2018 – Aug 2022
Bangalore, India

Bachelor of Engineering, Information Science and Engineering
Ramaiah Institute of Technology
First Class with Distinction

Publications

18 Sep 2023

Finding exoplanets using object detection
by S. R. Mani Sekhar, C. Tejas, **V. S. Navneet Kanna**, Aasees Kaur
Astrophysics and Space Science (Springer), 2023 [Paper] [Code]

Projects

FlashAttention 2 Triton Kernel (with Fused RoPE) [Code]

- Implemented the FlashAttention 2 forward pass from scratch in Triton, including online softmax, causal masking with early-stop, and fused Rotary Positional Embeddings (RoPE).
- Achieved **~145 TFLOP/s** (**~88%** of RTX 4090 peak) and matched and sometimes edges ahead of PyTorch's SDPA (FA2 backend) across sequence lengths 512–8192.
- **Fused RoPE** directly into the kernel to avoid extra HBM round-trips for Q/K and added causal early-stopping, which nearly doubled throughput at long sequences (**79 → 145** TFLOP/s at N=8192).
- Wrote two detailed technical blog posts with full derivations, worked examples, and benchmarks [Blog] .

dlgrad (Deep Learning Autograd Engine) [Code]

- Built a full reverse-mode automatic differentiation engine with PyTorch like API in Python plus custom C backend (**only one dependency**).
- Handwritten shape optimized kernels for **CPU** and **Apple Metal GPU**. Implemented full tensor stack, broadcasting, activations, and losses.
- Supports safetensors and pip installation.
- Implemented and trained **MLP, GAN, and GPT** architectures on a MacBook, generating original stories from the TinyStories dataset.

MSc Dissertation - ML Surrogate Models for Superconducting Flux Pump Simulation

- Built ML surrogate models to replace computationally expensive numerical simulations of superconducting flux pumps for fusion applications, where the numerical methods are too slow to be practical.
- Generated a training dataset from the numerical simulator and trained and compared five ML models for the prediction task.
- Showed the surrogate models reproduce the numerical results at a fraction of the runtime, supporting ML as a viable replacement. **Awarded A5 (Excellent)**.

Finding Exoplanets using Object Detection

- Used the YOLOv5 algorithm to detect planetary transit dips in light curve images from the TESS satellite, achieving **82% mAP** on test data.

fcount (High-Performance File System Counter)

- Developed a **command-line** tool in Rust that outperforms standard Linux utilities (ls, wc) for file system statistics.
- Achieved a **3.4x** speedup over GNU coreutils and **1.7x** over Go-based alternatives by bypassing redundant stat() syscalls and reading Linux d_type directory entries directly from the kernel buffer.
- Built a CI/CD pipeline using GitHub Actions to automatically test and release cross-platform binaries for Linux (x86_64) and macOS (Apple Silicon).

Blog

- Maintain a technical blog at navneetkanna.com/blog on GPU programming and deep learning internals.

Skills

Languages

- Python, C, C++, Rust

Deep Learning

- PyTorch, Triton, NumPy

Systems & Infrastructure

- CUDA, Apple Metal, Linux, Git, Docker, REST APIs