Aman Swar

India p.amanswar@gmail.com +91 7303166961 aman-swar-395826278 AmanSwar

Summary

AI Systems Engineer specializing in CUDA, PyTorch, and high-performance deep learning systems. Optimized CUDA kernels (BLAS, deep learning ops) achieving 80–90% of peak PyTorch performance, surpassing cuBLAS in some cases. Currently developing EANNS, a CUDA-accelerated vector database for billion-scale queries. Experience implemented a wide range of research papers spanning computer vision, NLP, and self-supervised learning, including IJEPA, MoCo, DINO, SimCLR, and transformer-based architectures. Experience in model optimizations, reducing inference latency and FLOPs by 40% on edge devices. Passionatye about low-level AI infrastructure, deep learning acceleration, and large-scale optimization.

Technical Skills

- **Programming Languages:** Python, C++, C, CUDA
- **AI/ML Frameworks:** PyTorch, TensorFlow, Hugging Face Transformers
- **GPU Computing:** CUDA Programming, cuBLAS, Parallel Computing
- **Machine Learning:** Deep Learning, Natural Language Processing, Computer Vision, Representation Learning
- **Optimization:** Model Quantization, Pruning, Distributed Training, OpenVINO, ONNX
- **Databases:** MySQL, FAISS, Vector Database
- **Computer Science Fundamentals:** Algorithms, Data Structures, Operating Systems, Computer Architecture and Organization, GPU Architecture, Parallel Computing

Experience

Undergraduate Researcher *10/2024 - Present*
SRM Institute of Science and Technology

- **Independently leading research and engineering** for an end-to-end AI system for automated diabetic retinopathy detection using Representation Learning. Responsible for all phases from algorithm selection and experimental design to implementation, optimization, and deployment planning.

Open Source

KernelLab – High-Performance CUDA Kernels for Deep Learning *02/2025 – Present*
(UNDER DEVELOPMENT) *KernelLab*

- **My open-source** CUDA optimized kernel library with optimizations like **shared memory utilization, memory coalescing, warp-level parallelism, and tensor core acceleration (WMMA API)**.
- Implemented high-performance **CUDA kernels** for deep learning operations including **Conv2D, Conv3D, SoftMax, Vector Add and ReLU, optimized with shared memory, warp-level parallelism, vectorized Execution and tensor cores, reaching 82%-90% of PyTorch implementation.**
- Implemented optimized Matrix Ops Kernels **(GEMM, Transpose, Reduction)** through **register blocking, coalesced memory access patterns, and warp tiling** techniques, **exceeding cuBLAS implementation** in matrices of sizes 256, 512, 1024 and reaching 67.4% of cuBLAS speed in bigger size matrices.
- Created accelerated image processing kernels **(Greyscale, Blur) utilizing vectorized multi-element processing and FMA optimizations**.
- UNDER DEVELOPMENT :- Self-attention and Flash attention

Efficient Approximate Nearest Neighbor Search (EANNS) — Vector Database *03/2025 – Present*
(UNDER DEVELOPMENT) *EANNS*

- **My open-source** high-performance **vector database** for large-scale similarity search using **C++/CUDA with Python bindings**, supporting billion-scale vector collections.
- Implemented **GPU-accelerated search algorithms** supporting multiple distance metrics (cosine, eu-

clidean, dot product) in CUDA.

- Designed and implemented efficient tensor-based storage with **metadata support** enabling **hybrid search capabilities (vector + metadata filtering)**.
- UNDER DEVELOPMENT :- efficient searching techniques and cache mechanism for hybrid storage

Projects

Deep Learning Pipeline for Diabetic Retinopathy Detection *10/2024 – 02/2025*
DR-detection

- Built an end-to-end deep learning pipeline for automated diabetic retinopathy diagnosis using self-supervised learning.
- Implemented and compared multiple **self-supervised learning papers including SimCLR, BYOL, DINOv2, iBOT, IJEPA**.
- Applied and customized various vision models such as **ViT, Swin Transformer, ConvNeXt** for different self-supervised techniques.
- Applied advanced CNN methods like **CBAM , Grade consistence , Gradient Reversal and custom loss function named OrdinalDomainLoss**
- Applied model compression techniques including **pruning (30% parameter reduction) and quantization (INT8)**, enabling deployment on resource-constrained devices.
- Integrated **explainable AI methods (Grad-CAM, Gradient based saliency maps, Layer-wise Relevance Propagation)** to visualize model attention and provide clinical interpretability.
- Optimized models with **OpenVINO** for deployment in low end devices with just CPUs.

SearchSphere – Multimodal Search Engine for Windows *01/2025 – 02/2025*
SearchSphere

- Engineered a **multimodal search engine** for Windows enabling natural language queries across documents and images with semantic understanding.
- Dramatically enhances Windows file search, **delivering speed improvements from 2.5x to 600x over standard search**.
- Utilized **FAISS** for efficient similarity search and implemented dual embedding pipelines **(MobileCLIP for images, BERT for text)** to support cross-modal queries.
- Designed a real-time indexing system supporting multiple file formats (.pdf, .docx, .txt, .jpg, .png) with automatic content extraction.

Education

SRM Institute of Science and Technology *2023 – 2027*
B.Tech in Computer Science with specialization in Artificial Intelligence and Machine Learning

- **Current GPA: 9.79/10**