AMAN SWAR

India | p.amanswar@gmail.com | +91 7303166961
linkedin.com/in/aman-swar | github.com/AmanSwar | amanswar.github.io

Summary

- AI Systems Engineer with expertise in CUDA programming, PyTorch optimization, and high-performance deep learning systems
- Creator of TorchSSL (Self-Supervised Learning Library), KernelLab (Optimized CUDA Kernels), and EANNS (Vector Database)
- Experience implementing and optimizing computer vision and NLP models for production environments
- Skilled in model optimization techniques achieving up to 60% reduction in memory requirements for edge deployment

Technical Skills

- **Programming Languages:** Python, C++, C, CUDA
- **AI/ML Frameworks:** PyTorch, TensorFlow, Hugging Face Transformers
- **GPU Computing:** CUDA Programming, cuBLAS, Parallel Computing, Tensor Core Optimization
- **Machine Learning:** Deep Learning, NLP, Computer Vision, Self-Supervised Learning, Explainable AI
- **Optimization:** Model Quantization, Pruning, Distributed Training, OpenVINO, ONNX Runtime
- **Databases:** MySQL, FAISS, Vector Databases, Similarity Search
- **Computer Science:** Algorithms, Data Structures, Operating Systems, Computer Architecture, GPU Architecture

Experience

Undergraduate Researcher
SRM Institute of Science and Technology, India

- Leading end-to-end research and development for automated diabetic retinopathy detection using representation learning techniques
- Developed RetinaSys, a state-of-the-art system for diabetic retinopathy detection optimized for edge devices, improving accessibility in underserved clinical settings (research paper submitted to journals)
- Engineered an AI-driven curriculum framework using large language models and retrieval-augmented generation to deliver personalized educational content
- Independently managing all project phases including algorithm selection, experimental design, implementation, optimization, and deployment planning

Open Source Projects

TorchSSL – High-Performance Self-Supervised Learning Library
Mar 2025 – Present

- Developed a PyTorch library for self-supervised learning that enables researchers to implement advanced techniques like SimCLR and DINO with minimal code
- Engineered optimized framework with support for popular SSL algorithms (SimCLR, MoCo, DINO, I-JEPA) and modern architectures (ViT, ConvNext)
- Created custom CUDA kernels for NT-Xent loss calculation, achieving 3-5× speedup over standard PyTorch implementation
- Implemented comprehensive SSL pipeline with integrated data loading, model training, and evaluation capabilities
- Added kNN and linear probe evaluation methods with latent space visualization tools
- Designed intuitive API allowing complete SSL training implementation in under 20 lines of code

KernelLab – High-Performance CUDA Kernels for Deep Learning
Feb 2025 – Present

- Built a CUDA-optimized kernel library utilizing shared memory, memory coalescing, warp-level parallelism, and tensor core acceleration (WMMA API)
- Implemented optimized kernels for core deep learning operations (2D/3D Convolution, SoftMax, Vector Addition, ReLU) achieving 82-90% of PyTorch performance
- Developed high-performance matrix operation kernels (GEMM, transpose, reduction) with register blocking and warp tiling techniques
- Achieved performance exceeding cuBLAS implementation for matrix sizes 256-1024 and reaching 67.4% of cuBLAS speed for larger matrices
- Created accelerated image processing kernels (grayscale conversion, blur filters) using vectorized processing and FMA optimizations
- Currently implementing self-attention and flash attention algorithms for transformer model acceleration

EANNS – Efficient Approximate Nearest Neighbor Search
Mar 2025 – Present

- Developing a high-performance vector database for large-scale similarity search using C++/CUDA with Python bindings
- Implemented GPU-accelerated search algorithms supporting multiple distance metrics (cosine, Euclidean, dot product)
- Designed efficient tensor-based storage architecture with metadata support enabling hybrid search capabilities
- Building optimized indexing and caching mechanisms for improved query performance in hybrid storage solutions

Projects

Deep Learning Pipeline for Diabetic Retinopathy Detection
Oct 2024 – Mar 2025

- Created an end-to-end deep learning pipeline for automated diabetic retinopathy diagnosis using self-supervised learning
- Implemented multiple state-of-the-art self-supervised methods (SimCLR, BYOL, DINOv2, iBOT, IJEPA)
- Adapted and customized advanced vision models including ViT, Swin Transformer, and ConvNeXt for medical imaging tasks
- Integrated attention mechanisms (CBAM), domain adaptation techniques, and developed a custom OrdinalDomainLoss function
- Achieved state-of-the-art performance with QWK (90.73%), AUC (90.85%), and F1 score (82.63%)
- Optimized model with OpenVINO, reducing RAM usage by 34.10% (FP16) and 60.07% (INT8) for efficient edge deployment
- Incorporated explainable AI methods (attention maps, integrated gradients, SHAP, Monte-Carlo dropout) for clinical interpretability

SearchSphere – Multi-modal Search Engine for Windows
Jan 2025 – Feb 2025

- Built a multi-modal search engine for Windows enabling natural language queries across documents and images
- Enhanced file search performance by 2.5-600× compared to standard Windows search capabilities
- Implemented FAISS for efficient similarity search with dual embedding pipelines (MobileCLIP for images, BERT for text)
- Designed real-time indexing system supporting multiple file formats (.pdf, .docx, .txt, .jpg, .png) with automatic content extraction

Education

B.Tech in Computer Science
Specialization in AI and Machine Learning
SRM Institute of Science and Technology

- **Current GPA:** 9.79/10