AMAN SWAR

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

Summary

- **High-Performance Computing for AI**: Proficiency in low-level kernel code using CUDA and Triton.
- **ML System Design & Frameworks**: Architected and built a modular self-supervised learning library in PyTorch (TorchSSL) and a distributed training library in JAX (DistJax) from the ground up.
- **Model Optimization & Deployment**: Experience in optimizing models for production, using techniques like quantization with OpenVINO to reduce memory footprint by 60% for edge deployment without loss of accuracy.

Technical Skills

- **GPU Computing & Performance:** C++, Python, CUDA, Triton, CUTLASS, CuTe, WMMA, NVIDIA Nsight Compute, GPU Architecture, Parallel Algorithms
- **ML Frameworks & Libraries:** PyTorch, JAX / Flax, OpenVINO, FAISS, Hugging Face
- **ML Concepts & Architectures:** Self-Supervised Learning, Distributed Training, Efficient ML (Quantization, Pruning), Computer Vision, NLP, Transformers, LLMs, Explainable AI (XAI)

Experience

Undergraduate Researcher **Oct 2024 – Present**
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) Pre-Print link.
- Engineered an AI-driven curriculum framework using large language models and retrieval-augmented generation to deliver personalized educational content.

Key Open Source Projects

KernelLab – High-Performance CUDA Kernels **Feb 2025 – Present**
github.com/AmanSwar/KernelLab

- Implemented optimized CUDA kernels for deep learning operations (Conv2D/3D, ReLU, RMSNorm, SoftMax, SwiGLU), BLAS operations (MatMul, Transpose, Reduction), and image processing (Grayscale, Blur).
- Implemented optimized Triton kernels for Deep Learning operations (softmax , Layer Norm , RoPE , SwiGLU , GeGLU and Flash attention) and BLAS operations (vector addition , Matrix Multiplication , Group Matrix Multiplication).
- Developed progressive optimization levels from naive implementations to highly-tuned kernels via extensive profiling using NVIDIA Nsight Compute CLI using memory coalescing, shared memory optimization, and advanced CUDA techniques.
- Built dual-precision support (FP32/FP16) with comprehensive performance analysis across different optimization levels.
- Benchmarked against industry-standard libraries (cuBLAS, cuDNN, PyTorch) achieving significant performance improvements over baseline implementations.

DistJax - Mini distributed training library in Jax **Aug 2025 – Present**
github.com/AmanSwar/DistJax

- Architected and developed DistJax, a comprehensive distributed training library in JAX and Flax, to simplify and scale deep learning models across multi-device environments.
- Implemented and benchmarked three core parallelism strategies: Data Parallelism (for data throughput), Tensor Parallelism (for large models), and Pipeline Parallelism (for deep models).
- Engineered advanced asynchronous communication primitives for Tensor Parallelism using JAX's ppermute, effectively hiding communication latency and improving hardware utilization.
- Authored end-to-end model implementations, including a fully tensor-parallel Transformer, to validate the library's effectiveness and provide practical usage examples for researchers.

TorchSSL – Self-Supervised Learning Library **Mar 2025 – Present**

github.com/AmanSwar/TorchSSL

- Developed a high-performance, modular PyTorch library for Self-Supervised Learning (SSL) implementing SimCLR, MoCo, DINO, and I-JEPA frameworks.
- Engineered custom, fused Triton kernels for NT-Xent and InfoNCE loss (and many more coming up..) functions, achieving significant speedups over standard PyTorch implementations.
- Designed a flexible and extensible framework with support for various backbones (e.g., ConvNeXt, ResNet), comprehensive evaluation suites (kNN, Linear Probing), and integrated visualization tools (WandB, PCA/t-SNE).
- Created a streamlined data loading and augmentation pipeline, enabling efficient training on large-scale, unlabeled image datasets.

Selected Projects

Diabetic Retinopathy Detection Pipeline **Oct 2024 – Mar 2025**

github.com/AmanSwar/DR-detection

- 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 **Jan 2025 – Feb 2025**

github.com/AmanSwar/SearchSphere

- Built a multi-modal search engine for Windows enabling natural language queries across documents and images.
- Enhanced file search performance by 2.5-100× 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 (AI & ML Specialization) **2023 – 2027**

SRM Institute of Science and Technology

- **Current GPA:** 9.7/10