AMAN SWAR

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

Summary

AI Systems Engineer specializing in high-performance ML infrastructure with full-stack capabilities from research to metal. Built production-grade optimization libraries (TorchPP: 9.25x inference speedup, DistJax: multi-paradigm distributed training) using custom CUDA/Triton kernels. Experienced across the entire ML pipeline: novel architecture design for edge deployment, GPU kernel optimization, quantization (60% memory reduction), and distributed training systems. Combine low-level systems programming with end-to-end ML research and deployment.

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, TensorRT-LLM, vLLM, FlashInfer
- **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 – April 2025**
SRM Institute of Science and Technology, India

- Lead 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 introducing a novel convolution based architecture optimized for edge devices, improving accessibility in under served clinical settings (research paper submitted to journals) Pre-Print link.
- Designed an AI-driven curriculum framework using large language models and retrieval-augmented generation to deliver personalized educational content.

Key Open Source Projects (Projects I am actively working on)

TorchPP : add-ons for Pytorch for inference and distributed training **Nov 2025 - Present**
github.com/AmanSwar/TorchPlusPlus

- Built a **high-performance PyTorch extension with custom CUDA/Triton kernels** (fused Linear+Activation, LayerNorm/RMSNorm, RoPE, KV-Cache, Flash-style Attention variants).
- Achieved **9.39×** throughput and 9.25× latency reduction on Qwen-0.6B; **3.18×** latency & throughput improvement on a TTS model using my TorchPP.
- Implemented **speculative decoding, Multi-Query/Grouper/Cross/Sliding-Window attention, and an lightweight inference engine** optimized for all kind of transformer based models.
- Developed a distributed training abstraction supporting **DDP, FSDP, hybrid parallelism, AMP, gradient accumulation, and checkpointing**.

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.

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.

FastQwen3 - Qwen3 0.6B but faster **Sept 2025**
https://github.com/AmanSwar/FastQwen3

- Optimized Qwen3 0.6B parameter model in fp16 to run faster in consumer GPU , **achieving 9.25x ++ inference speedup over huggingface baseline**
- Implemented **KV Cache , fused RMSNorm , RoPE and custom Flash attention kernel to support Grouped Query Attention** in CUDA as well as in Triton
- **Reduced 600-token inference time from 440s to 48s (saving 6.5 minutes per request) and averaging 4.83x speedup for less than 600 tokens and 13.85x speedup for greater than 600 tokens**

Selected Projects

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.

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) without any loss in accuracy for efficient edge deployment.**
- Incorporated **explainable AI methods (attention maps, integrated gradients, SHAP, Monte-Carlo dropout)** for clinical interpretability.

Education

B.Tech in Computer Science (AI & ML Specialization) **2023 – 2027**
SRM Institute of Science and Technology

- **Current GPA:** 9.7/10