

Srijan Upadhyay

Distributed ML Systems · LLM Infrastructure · Computer Vision

Pune, India | srjnupadhyay@gmail.com | linkedin.com/in/srijan-upadhyay-a7293a240 | github.com/CodersAcademy006

SUMMARY

ML systems engineer with deep focus on distributed training infrastructure, LLM inference optimization, and scalable computer vision pipelines. Contributor to major ML frameworks with 4 merged PRs across TensorFlow, HuggingFace Transformers, and Megatron-LM, and an active code reviewer on TensorFlow core. Founder of ScoreOne, an AI-backed ed-tech startup. Pursuing B.Tech in Data Science at SPPU (graduating 2028).

TECHNICAL SKILLS

ML Systems: Distributed Training, FSDP, MoE Architectures, NCCL, Async Scheduling, vLLM, TensorRT

Frameworks: PyTorch, TensorFlow, HuggingFace Transformers, Megatron-LM, TransformerEngine, NeMo

Infra / CUDA: CUDA, Numba-CUDA, Triton, ONNX Runtime, TensorRT, H100 cluster management

CV / Retrieval: InsightFace, OpenCV, FAISS, Qdrant, FastAPI, ONNX Runtime

Languages: Python (primary), TypeScript, Java, R, JavaScript

Tools: Git, Docker, Railway, Streamlit, Colab/T4, REST APIs, Flask, Back4App

EXPERIENCE

Founder & Lead Engineer — ScoreOne (Ed-Tech Startup, scoreone.in)

Remote | Apr 2026 – Present

- Building an AI-backed ed-tech platform for JEE/NEET exam prep (scoreone.in): a full-stack mock-test and online-proctoring product on a Back4App Cloud Code backend, TypeScript/React frontend, and a parallel Python reference backend.
- Shipped NTA-parity scoring, checkpoint-based attempt integrity, and server-side anti-cheat: phone split-screen detection plus server-enforced 3-violation Unfair-Means auto-submit with full session teardown (no client trust on cheat state).
- Owned production reliability and security end-to-end — removed hardcoded credentials, scoped production dependency audits, and added ESLint CI guards on the deployed Cloud Code backend; 180+ merged PRs to date.

OPEN SOURCE CONTRIBUTIONS

Authored — Merged into top ML frameworks

Megatron-LM #2571 — Merged Dec 2025

- Fixed distributed initialization deadlock from incorrect world-size derivation via `torch.cuda.device_count()` instead of `torchrun` env vars (`WORLD_SIZE`, `RANK`, `LOCAL_RANK`); resolved process group failures across multi-GPU (2+ node) training runs.

Megatron-LM #2765 — Merged Feb 2026

- Fixed numerical instability in MoE auxiliary loss by upcasting sigmoid routing scores to fp32; prevented precision loss during BF16/FP16 training with large expert counts (32+ experts), aligning behavior with the existing softmax path.

TensorFlow #102094 — Merged Oct 2025

- Fixed Google Play validation error for 16KB memory page sizes on Android 15+ by adding `-Wl,-z,max-page-size=16384` linker flag to the `:native_stable_framework_only` JNI build target; unblocked app store submission for all dependent JNI libraries.

HuggingFace Transformers #43047 — Merged Apr 2026

- Fixed FSDP and model parallelism crash in `PeVideo` and `PeAudioVideo` by setting `_no_split_modules=["TimmWrapperForImageClassification"]`; prevented accelerate from layer-splitting the timm vision encoder across devices, eliminating `RuntimeError` in distributed training; `test_model_parallelism` now passing.

Code Review — TensorFlow core (peer review of security & correctness fixes)

- Maintainer-requested reviewer on tensorflow/tensorflow #118744 (tf.bitcast / tf.cast eager-vs-XLA divergence); provided detailed change-requesting review.
- Reviewed security and correctness PRs: OOB write in `MaxPoolGradWithArgmax` GPU kernel / CWE-787 (#115326), unbounded recursion in Variant deserialization (#118269), Grappler `ArgMin/ArgMax` rewrite over non-monotonic ops (#116709), and runtime rank invariants in `StringNGrams` (#118868).

PROJECTS

Midwicket — Open-Source Cricket Intelligence SDK (PyPI)

- Authored an open-source, local-first ball-by-ball cricket analytics SDK (`pip install midwicket`, MIT) over 20,000+ matches / 9.1M+ deliveries / 25+ years, built on DuckDB + PyArrow + FastAPI — no cloud required.
- Shipped win-probability prediction with a versioned model registry (semantic versioning + rollback), feature registry, reproducibility/leaderboard runner, Redis-backed rate limiting, full mypy typing, and CI; resolved 100+ tracked correctness/UX defects across 35+ merged PRs.

Academic projects — SPPU Department of Technology | Nov 2025 – Mar 2026

FotoOwlAI — Cloud-Native Face Recognition Attendance System

- Architected fully cloud-native pipeline across 4 services: Colab T4 inference (InsightFace buffalo_l + glintr100) → Qdrant Cloud (512-dim embeddings) → Railway (Telegram bot) → Google Drive; designed for multi-department SPPU deployment.
- Implemented `norm_crop()` 112×112 preprocessing + `get_feat().flatten()` embedding pipeline; collection-level re-indexing on model switch with zero downtime.
- Integrated DPDP Act 2023 compliance (biometric consent + data handling policy) into system design from day 1; resolved critical Qdrant SDK incompatibility (`≥v1.7 .search() → .query_points()`) and InsightFace model-path bugs.

Geospatial Analytics Viewer + Financial Dashboards

- Built an interactive geospatial analytics platform (GeoPandas + Streamlit) plus structured ETL pipelines and financial dashboards (Python + Matplotlib) for enterprise-scale reporting.

EDUCATION

B.Tech in Data Science — Savitribai Phule Pune University (SPPU)

Pune, India | Oct 2024 – Present

- GPA: 8.5/10 · Relevant coursework: Machine Learning, Distributed Systems, Computer Vision, Statistical Modeling, Algorithms