Tanmay Sutar

Atlanta, GA, USA

+1 (404) 423-9768 | tsutar3@gatech.edu | https://www.linkedin.com/in/tanmaysutar2109/
https://github.com/Tanny2109

Summary

New-grad Machine Learning Engineer specialized in multi-GPU systems and 3D computer vision projects. Architected a multi-GPU training pipeline for LLaMA, achieving high accuracy in empathy detection, and improved GPU load balancing for enhanced processing throughput. Seeking to leverage expertise in machine learning to improve workflow efficiency and real-time data processing in target job roles.

EDUCATION

Georgia Institute of Technology **Aug 2023 - May 2025**
Master's, Computer Science - ML Specialisation *Atlanta, GA, USA*

IIT Guwahati **Aug 2019 - Jun 2023**
Biotechnology *Guwahati, India*

EXPERIENCE

Social Wellbeing **Jan 2025 - May 2025**
Machine Learning Research Internship - Red-teaming Llama *Atlanta, USA*

- Architected multi-GPU training pipeline for jailbreaking LLaMA 3.1-8B with SFT (Supervised Finetuning) and LoRA fine-tuning, achieving 0.98 ATS score on toxicity detection benchmarks (HarmBench, LLM-as-judge).
- Built a distributed inference system with dynamic GPU load balancing across 4 Nvidia L40S GPUs, processing 5K+ conversations with 4x throughput improvement using async batch processing with ollama.
- Developed production-ready AutoGen integration with multi-threaded and multi-gpu conversation simulation, implementing fault-tolerant error handling and GPU resource optimization for 24/7 deployment.

Bio-MIB **Aug 2024 - Dec 2024**
ML Research Internship - 3D Computer Vision *Atlanta, USA*

- Engineered 3D brain tumor segmentation pipeline using 3D U-Net and VMNet architectures; achieved Dice coefficient of 0.83 on BraTS dataset with custom data augmentation strategies.
- Implemented distributed PyTorch training with Luigi workflow orchestration on CUDA 12 infrastructure, improving compute efficiency by 70% and reducing training time from 48 to 14 hours.

CNRS - Applied ML Lab **Jan 2024 - May 2024**
Computer Vision Internship - Video Analytics *Metz, France*

- Fine-tuned YOLOv8 object detection model for judo match analysis, achieving mAP@0.5 = 0.92 and enabling automated highlight generation with 95% accuracy.
- Integrated YOLOv8 with LabelStudio for active learning annotation pipeline, reducing manual labeling effort by 40% and accelerating dataset creation for sports analytics.
- Deployed real-time video processing system with OpenCV and FFmpeg, processing 1080p video streams at 30 FPS for live match analysis.

PROJECTS

Large-Scale AI Bias Detection Platform **May 2024 - Aug 2024**
Atlanta, USA

- Architected end-to-end ML pipeline processing 10K+ Reddit posts, deploying dual-model sentiment analysis (GoEmotions RoBERTa + VADER)
- Optimized inference performance by 10x through ONNX runtime integration, reducing analysis time from 48 hours to 4.8 hours.
- Delivered actionable bias insights that influenced content moderation policies for Reddit.

Audio Recognition Model Optimization **Jan 2025 - May 2025**
Atlanta, USA

- Fine-tuned Whisper and Wav2Vec models for specialized audio recognition tasks, improving accuracy by 17% over baseline implementations.
- Implemented custom data augmentation pipeline resulting in 35% reduction in error rates for noisy audio environments.

Skills

- **Relevant Skils**: Python, PyTorch, LLMs, SQL, Spark, CUDA, Object Oriented Programming, AWS, GCP, Transformers, Applied Research