Aakash Kaushik

Software Engineer

kaushikaakash7539@gmail.com +917575885094 Aakash-kaushik kaushikaakash7539

Summary

Highly productive Full-Stack AI Engineer with 4+ years of experience driving innovation at Tune AI and contributing significantly to open-source machine learning. Authored 5.5M+ lines of code across 100+ projects, specializing in building scalable AI solutions and cloud infrastructure. Expertise in machine learning, distributed systems, and full-stack development with Python, Go, and C++. Proven ability to deliver high-impact projects in NLP, computer vision, and MLOps, leveraging emerging technologies like Kubernetes, Docker, and serverless computing for enhanced performance and reliability. Passionate about open source and community contributions, with demonstrated leadership in projects like mlpack (116K+ lines of code, 13 projects).

Skills

Python, Golang, C++, GCP, AWS, Azure, Kubernetes, Docker, TensorFlow, PyTorch, OpenVINO, CI/CD, Generative AI, Machine Learning Systems, NLP, Computer Vision, MLOps, Serverless Computing, REST APIs, Git

Experience

Software Engineer *Oct 2020 – Present*
Tune AI

- Led development of core services for generative AI systems, resulting in a 20% reduction in latency and a 15% improvement in throughput.
- Architected and deployed sidecar servers for multi-cloud instance management, increasing deployment speed by 25%.
- Optimized data handling processes for object stores, achieving a 10% reduction in storage costs.
- Contributed to the design and implementation of multiple cloud-native applications using Kubernetes, Docker, and serverless technologies.
- Developed and maintained APIs and backend services for key products, serving 100K+ users.

AI/ML Intern *Nov 2021 – Jun 2022*
OptimEyes.ai

- Developed AI-powered risk assessment models for cybersecurity, resulting in a 30% improvement in threat detection accuracy.
- Automated client cloud risk analysis and optimization processes, saving 20+ hours per week.
- Implemented machine learning algorithms for enhanced risk assessment, reducing false positives by 15%.

Developer - mlpack *May 2021 – Aug 2021*
Google Summer of Code 2021

- Created MobileNetV1 and a ResNet model builder, simplifying model construction and improving user experience.
- Integrated pre-trained model weights, reducing model training time by 40%.
- Contributed 2K+ lines of code to mlpack, including bug fixes and feature implementations.

Deep Learning Engineer *May 2020 – Aug 2020*
Mavoix Solutions

- Built text recognition and image classification models from scratch, achieving 95% accuracy on benchmark datasets.
- Improved model performance by 10% through ensemble techniques and component optimization.
- Developed and deployed Flask-based APIs for seamless model integration, reducing integration time by 50%.

Projects

Studio
Tune AI

- Helped Develop a scalable backend system for fine-tuning, multi-modal models, including features like BYOC support, integrations with Weights & Biases and LangChain, and token limits. (500K+ lines of code, 250+ commits)

LLM Proxy Server
Tune AI

- Developed a highly performant proxy server to route API requests, handling 1M+ requests per day. Implemented support for various language models, token limits, and caching strategies. (2.5M+ lines of code, 250+ commits)

Infra Engine
Tune AI

- Built a cloud infrastructure management engine using Pulumi and Kubernetes, automating infrastructure provisioning and deployment across GCP and AWS. (30K+ lines of code, 150+ commits)

Relics Server
Tune AI

- Developed a backend system for managing files and logs, ensuring scalability and fault tolerance. Implemented support for cloud storage backends like S3, Azure Blob Storage, and Google Cloud Storage. (40K+ lines of code, 100+ commits)

SDXL Triton
Tune AI

- Implemented Stable Diffusion XL support on Triton Inference Server, enabling TPU deployment and optimizing image generation. (20K+ lines of code)

Models
mlpack

- Revamped documentation, built a CI/CD pipeline, and contributed MobileNetV1 and ResNet integrations. (100K+ lines of code, 90 commits)

Technologies

Languages: C++, C, SQL, Python, Go

Technologies: SQL Server, GCP, AWS, Azure, Kubernetes, Docker, TensorFlow, PyTorch, OpenVINO, CI/CD, Generative AI, Machine Learning Systems, NLP, Computer Vision, MLOps, Serverless Computing, REST APIs, Git