Bijon Setyawan Raya

bijonsetyawan@gmail.com github@bsraya · https://bsraya.com

Experience

Senior Machine Learning Engineer October 2025 – Now
IKG Team Ltd. *Taipei, Taiwan*

- Design and lead the design and development of a Content Moderation platform leveraging Guardrails for messaging services, achieving up to 80% latency reduction and nearly 90% decrease in total token consumption over time.

Machine Learning Engineer November 2024 – October 2025
Cathay United Bank *Taipei, Taiwan*

- Designed, engineered, and deployed Guardrails on AWS to consolidate and safeguard enterprise access to LLMs which eliminates 99% of unsafe interactions and reduced cut response latency up to 90% from previous version.
- Performed deep performance analysis and profiling of high-performance LLM inference (vLLM) and fine-tuning (LLaMA-Factory) frameworks to quantify the architectural trade-offs between bare-metal and virtualized NVIDIA GPU environments (MIG, time-slicing) on RHEL OS servers.

Intermediate Fullstack Developer October 2022 – March 2024
Faria Education Group *Taipei, Taiwan*

- Reduced file export time by 98.33% and improved user feed search speed and accuracy by 40% through performance optimizations and algorithmic improvements.
- Refactored 25% of legacy jQuery codebase to Stimulus.js improving maintainability and reducing bug rates by 30%.

Technical Skills

Languages: C/C++, Golang, JavaScript/TypeScript, Python*.
Web Development: Astro.js, FastAPI*, Next.js, Solid.js.
Deep Learning & Machine Learning: CUDA, cuML, PyTorch*, Scikit-learn.
Database: DuckDB, ElasticSearch, MySQL, PostgreSQL, SQLite, Redis.
Utilities: Docker*, Git, Kubernetes*, Linux/Unix, Podman*, uv.
Cloud Services: Amazon Web Service*, Google Cloud Provider.

Projects

MLOps | *FastAPI, Scikit-Learn, Redis, cuML, Min.io, MLFlow, Optuna*

- Automated and reduced training time by 50%, enabling seamless model deployment, effectively bridging the gap between data scientists and DevOps engineers.

Scikit-Learn in C++ | *C++, eigen, vcpkg*

- Built a C++ implementation of Scikit-learn with CUDA-optimized regression algorithms, achieving 10x speedup over Python implementations for large-scale datasets.

Image Search Engine | *Docker, PostgreSQL, FastAPI, PyTorch, ROCm, uv*

- Reducing image search latency by 95% using ROCm and reduce storage usage by 87% using PCA.

Music Recommendation System | *Numpy, Pandas, Python, scikit-learn*

- Leverages Spotify API data and Non-negative Matrix Factorization (NMF) to generate personalized music recommendations, effectively discovering new tracks aligned with user preferences.

Schedulearn | *Docker, FastAPI, Horovod, Python, SQLite*

- A lightweight distributed deep learning scheduling system that reduces make span by 50% and increased throughput by 70% across different servers and GPUs.

Education

National Tsing Hua University Hsinchu, Taiwan
Master of Science in Computer Science *January 2023*

National Tsing Hua University Hsinchu, Taiwan
Bachelor of Science in Computer Science *January 2021*