Mu Yang 楊慕

SENIOR SOFTWARE ENGINEER & TECH LEAD

☎ (+886) 920-890-202 | ✉ emfomy@gmail.com | ⌂ muyang.pro | 📍 Taipei, Taiwan | emfomy | in emfomy

Summary

Senior software engineer with a research background in computer science and applied mathematics. Specialized in building scalable backend systems with Go, and deploying GPT-based applications using LoRA and RAG in production. Experienced in infrastructure-as-code, container orchestration, and performance tuning in distributed environments.

Skills

Programming	Go, C/C++, Python, TypeScript
Machine Learning	PyTorch, TensorFlow, LLM, Transformers, LoRA, RAG, LangChain
Frameworks & Libraries	Gorilla Mux, gRPC, GORM, MPI/OpenMP, LAPACK/MKL, CUDA
Databases	PostgreSQL, Redis, MinIO
DevOps	Docker, Kubernetes, Helm, GitLab CI, ArgoCD, CDK8s, Terraform, AWS, GCP
Testing & Monitoring	Go testing, GoConvey, Mockgen, Prometheus, Grafana
API & Protocols	RESTful API, OpenAPI, Server-Sent Events (SSE)
Languages	Chinese (Native), English (Fluent; TOEFL 96)

Work Experience

Taiwan AI Labs

Taipei, Taiwan

SENIOR SOFTWARE ENGINEER & TECH LEAD

Feb. 2021 – Present

- **FedGPT**: On-premises GPT system integrating custom LoRA, RAG, and multi-agent architecture.
  - Developed a real-time GPT chat platform featuring streaming, RAG, and multi-agent workflows for enterprise use cases in private environments.
  - Managed production deployment of large-scale LLMs and other ML models, optimizing resource utilization and implementing dynamic model enablement controls.
  - Designed infrastructure-as-code (IaC) solutions with CDK8s and built CI/CD pipelines adopted by 6+ engineering teams (~30 developers), enabling reproducible deployments and efficient team workflows.
  - Led an engineering team of 10, overseeing architecture and implementation.
- **Miin**: AI-driven platform aggregating multi-platform insights to reduce social bias and polarization.
  - Implemented high-performance follow and feed systems optimized for future scalability and high fanout patterns.
  - Designed a modular file upload system with presigned URLs, CDN support, and auto-generated responsive images.
- **Infodemic**: AI-powered platform system detecting coordinated disinformation and narrative manipulation.
  - Led an engineering team of 8, overseeing architecture and implementation.
  - Built and maintained APIs and data pipelines handling billions of social comments and hundreds of millions of posts and articles.
  - Designed data models and indexing strategies to support fast content retrieval and precomputed analytics for web presentation.

CKIP Lab, Institute of Information Science, Academia Sinica

Taipei, Taiwan

RESEARCH ASSISTANT

July 2017 – Jan. 2021

- Supervisor: Dr. Wei-Yun Ma
- Conducted research on natural language processing and computational linguistics.
- **CKIP Transformers**: An end-to-end NLP toolkit with transformer models for Traditional Chinese.
  - Trained transformer models (ALBERT, BERT, GPT-2) for Traditional Chinese on a national supercomputing cluster.
  - Fine-tuned task-specific models for segmentation, POS tagging, and NER.
  - Released to HuggingFace with 240k+ monthly downloads.

Thomas J. Watson Research Center, IBM Corporation

Yorktown Heights, NY, U.S.A

INTERNSHIP

July 2015 – Aug. 2015

- Supervisor: Dr. I-Hsin Chung
- Researched high-performance computing on hybrid CPU-GPU systems using IBM supercomputers.
- Published a peer-reviewed paper on parallel feature selection.

Education

National Taiwan University *Taipei, Taiwan*

MASTER OF SCIENCE IN APPLIED MATHEMATICAL SCIENCES *Sept. 2015 – June 2017*

- Supervisor: Prof. Weichung Wang
- Conducted research of high-performance parallel computing on hybrid CPU-GPU structures.

National Taiwan University *Taipei, Taiwan*

BACHELOR OF SCIENCE IN MATHEMATICS *Sept. 2011 – June 2015*

Publications

12th Language Resources and Evaluation Conference (LREC'20) *Marsélle, France*

MU YANG, CHI-YEN CHEN, YI-HUI LEE, QIAN-HUI ZENG, WEI-YUN MA, CHEN-YANG SHIH, WEI-JHIH CHEN *May 2020*

Headword-Oriented Entity Linking: A New Entity Linking Task with Dataset and Baseline

13th IEEE International Conference on Semantic Computing (ICSC'19) *Newport Beach, CA, U.S.A*

JHIH-SHENG FAN, **MU YANG**, PENG-HSUAN LI, WEI-YUN MA *Jan. – Feb. 2019*

HWE: Word Embedding with Heterogeneous Features

16th IEEE International Conference on Computer and Information Technology (CIT'16) *Yanuca Island, Fiji*

MU YANG, RAY-BING CHEN, I-HSIN CHUNG, WEICHUNG WANG *Dec. 2016*

Particle Swarm Stepwise Algorithm (PaSS) on Multicore Hybrid CPU-GPU Clusters

Master's Thesis, National Taiwan University *Taipei, Taiwan*

MU YANG, (ADVISOR: WEICHUNG WANG) *July 2017*

Highly Scalable Parallelism of Integrated Randomized Singular Value Decomposition with Big Data Applications