Yunmo Koo

AI System Engineer (and all-rounder)

mpbb03@gmail.com +82 (010) 2311-9945 ***yunmorning.me*** *yunmokoo* *kooyunmo*

I am a Machine Learning Systems Engineer with a robust blend of academic achievements and industry experience. Over the past five years, I've built an end-to-end generative AI platform for training and deploying LLMs, playing a key role as a founding team member of a successful Series-A startup. My expertise spans the full lifecycle of LLM systems, from conception to real-world application.

Experience

FriendliAI *Feb 2021 - present*

- As a founding team member, I led the initial product development of PeriFlow, a platform to run distributed training jobs on any cloud (e.g., AWS, Azure, GCP), handling every aspect of fault tolerance and resource management.
  - Successfully trained an LLM model (FAI-13B) with PeriFlow and published the model a year ahead of Meta's Llama 2.
- Developed Friendli Engine, the core inference serving engine for higher throughput and lower latency.
  - Designed and implemented innovative speculative decoding techniques, achieving 2~3x latency improvement.
  - Implemented and optimized CUDA kernels required for efficient computation for various LLMs/LMMs.
  - Implemented various LLMs/LMMs inference supports, including Llama 4 and Gemma 3.
- Managed the full spectrum of system development and management, from setting up cloud infrastructure to designing and implementing key microservices of MLOps system including authentication, registry, training, deployment, and monitoring.
- Developed user interfaces, including SDK, CLI, and web frontend, and managed the official documentation site.
- Contributed to popular open-source frameworks to build RAG (Retrieval Augmented Generation)-based LLM applications such as LangChain and LlamaIndex.
- Managed collaborations with cloud providers, including publishing products on AWS and Azure marketplaces and participating in the AWS ISV program.
- I am much more than just an engineer; set up sales channels in the US market and have actively driven business development. My experience includes identifying market needs, acquiring customers through targeted campaigns, and cultivating lasting business relationships.
- Actively presented product demonstrations to B2B clients and at global events (AWS re:Invent, NVIDIA GTC, ICML), showcasing technical prowess and business acumen.

Software Platform Lab *@Seoul National University* *Aug 2020 - Aug 2022*

- Optimized deep learning computation graphs for enhanced performance.
  - ***Terra: Imperative-Symbolic Co-Execution of Imperative Deep Learning Programs***, Taebum Kim, Eunji Jeong, Geon-Woo Kim, **Yunmo Koo**, Sehoon Kim, Gyeongin Yu, Byung-Gon Chun, Advances in Neural Information Processing Systems 34 (**NeurIPS 2021**)
- Engineered cost-effective distributed training job orchestration tool across multiple cloud platforms.
  - ***Cost-Efficient Machine Learning Training on Preemptible Cloud Clusters***, **Yunmo Koo**, Master Thesis of Seoul National University Graduate School

Education

Seoul National University *Aug 2020 - Aug 2022*

- M.S. in Computer Science and Engineering

Seoul National University *March2014 - Aug 2020*

- B.S. in Computer Science and Engineering (double major)
- B.S. in Korean History
- The period includes two years of military service as a KATUSA (Korean Augmentation to the United States Army)

Skills

Language	Python, C++, CUDA, TypeScript, Rust
Framework	PyTorch, FastAPI, Django, NextJS
Tool	Kubernetes, Argo CD, Jenkins, Kafka, ElasticSearch, Prometheus, GraphQL
Cloud	AWS, Azure, GCP, CoreWeave, Nebius
Methodology	Machine Learning System, LLMOps, Multi-Cloud

Teaching

- **Principles and Practices of Software Development (Seoul National University, Fall 20)**
  - Served as a Teaching Assistant, conducting practical sessions on React, Django, CI/CD, and design patterns, and facilitated design meetings for student projects.