Xi GUO

im.guoxi@gmail.com — +86 13720367764 — www.linkedin.com/in/GUOXi

EDUCATION

Huazhong University of Science and Technology, Wuhan, China — September 2019 — June 2023
B.E. in Computer Science and Technology — Cumulative GPA: 3.74/4.00
Selected Courses: Data Structure, Operating Systems, Computer Architecture, Computer Network, Database Systems

PROFESSIONAL EXPERIENCE

Pinduoduo (NASDQ: PDD, parent company of TEMU) — Shanghai, China
Machine Learning Infrastructure Engineer (Training Frameworks) — July 2023 — November 2025

Generative Recommendation Models

- Implemented HSTU-based recommendation models with PyTorch + TorchRec, covering full training workflow (DDP + DMP: data loading, batching, training, parameter updates, checkpointing).
- Trained models on large scale of 300M+ DAU samples across recall, search CVR, and CTR scenarios; produced embeddings for downstream offline models.
- Developed high-throughput Torch Dataset Ops (C++/Python) for multi-format HDFS samples (TXT/ORC/Parquet), enabling multi-process parallel reading with CPU–GPU overlap, improving training speed by 40%+.

Training Framework Migration

- Migrated large-scale models from TensorFlow v1 + custom sparse parameter servers to PyTorch/TorchRec
- Scaled training on multi-node H100 clusters with 1B+ feature IDs, enhancing inter-node RDMA communication by optimizing NCCL Net plugins.
- Combined pretrained static sparse tables with TorchRec Embedding Collection, reducing GPU memory usage by 50%+.
- Achieved 30%+ higher GPU utilization and 15%+ faster training; after training 30 days' user data, AUC aligned with baseline ($\Delta \leq 0.0005$).

Semi-Synchronous Optimizer Development

- Extended SyncReplicaOptimizer(TF1.0) with Global Batch Aggregation (NeurIPS 2022) semi-synchronous mechanism.
- Implemented token-based gradient admission and staleness control, balancing updates and discarding stale gradients.
- In CVR tasks, improved AUC by 0.3%, achieved 60%+ speedup over native SyncReplicaOptimizer, with efficiency reaching 85% of async training.
- Supported partial parameter semi-sync updates (e.g. BatchNorm statistics) via TF graph modifications.

Distributed Training Coordination System

- Designed barrier synchronization and file distribution across training pods using Zookeeper + RPC.
- Implemented leader election with ZK; master node broadcasts server address and distributes configs via HTTP.

Machine Learning Platform Backend

- Contributed to the ML platform for rec/search engineers, supporting job submission, scheduling, and monitoring.
- Developed and maintained RESTful APIs (Python/Java) for job management, resource prioritization, user access control.
- Unified external API interfaces by merging dual services, implementing proxy forwarding with full consistency validation.

Sensetime — Shanghai, China
Backend Engineer Intern — July 2022 — September 2022

- Contributed to the development of SenseParrots Enterprise, a B2B deep learning platform covering end-to-end workflows from data management to model training.
- Arranged and optimized training task scheduling and deployment using Kubernetes and DockerHub, improving resource utilization and job reliability.
- Developed and maintained backend APIs for multiple microservices using Golang (Gin framework), enhancing system scalability and maintainability.

Pivot Studio — Wuhan, China
Backend Developer — October 2021 — July 2022

- Designed and maintained backend service of HUST-Hole, an anonymous campus discussion platform for Huazhong University of Science and Technology with 2,000+ daily active users (DAU).
- Built RESTful APIs and optimized database queries to improve response latency and platform stability.

SKILLS

C/C++, Python, GoLang, Java, Triton, CUDA, Tensorflow, PyTorch, Torchrec, Unix&Linux, Vim, Git

RESEARCH INTERESTS

Machine Learning Systems, ML and Systems Co-design, Distributed Training Framworks, High Performance Computing