Xingfu Yang, PhD

U.S. Permanent Resident | 303 523 8836 | xingfu@xyang.me | LinkedIn | GitHub | Los Angeles, CA

Highlights

- Accomplished **Founding Data Scientist** with extensive startup experience, spearheading the development of innovative data products from ideation to scalable profitable businesses.
- Hardware-aware AI Researcher with a solid track record of developing innovative solutions and conducting open-ended research to solve real-world problems, mindful of hardware characteristics and algorithm constraints.
- Abreast of latest algorithmic advances, experienced in refining scaling laws and composing different parallelism strategies when training massive datasets on supercomputers with resource accounting.
- Exemplifies the mindset of exploration vs. exploitation; proficient in high-level system design to hands-on development of quality distributed deep learning systems with 6 years of industrial experience, highlighting project scoping and problem solving at all layers of the stack.

Work Experience

Lucid Intel Santa Monica, CA

Chief Data Scientist *Jun 2023 – Present*

- Leading a team developing and delivering core data products, including lead acquisition, non-credit underwriting, and high-frequency fraud prevention, leveraging unsupervised and superwised deep learning on 40M identities and billions of records.
- Spearheading an AutoML-based platform to accommodate various client demands on top of off-the-shelf products through project scoping, customized modeling and continuous iteration.
- Co-created a new business division – an advanced aggregated data analytics platform offering managed agency of record services for major clients.

Leap Theory Los Angeles, CA

Data Scientist *Oct 2019 – Jun 2023*

- Orchestrated non-credit transformer-based aggregated models for InsightEngine to serve key lending underwriting metrics in production.
- Achieved a 55% reduction in unnecessary traffic, a 70% increase in accept ratio, and a 2-10 percentage point decrease in first payment default.

Independent Projects

MiniLM: a minimal JAX incarnation of full life cycle modern LM from scratch

- A TPU Research Cloud Project featuring multi-device sharded pretraining of a Mixture-of-Experts foundation model using the OpenWebText dataset with data and tensor parallelism.
- Post-training: supervised fine-tuning, reward modeling, followed by RLHF with PPO, DPO, and GRPO.

RelationalLearning: Scaling up GNNs with GraphStore and FeatureStore via Remote Backends

- Trained a GraphSAGE model on a 100 million nodes and 1.6 billion edges graph DB backing dataset

Expertise

Interests: Deep Learning Frontier ✖ Business: Graph & Relational Learning, GenAI, Reinforcement Learning

ML Frameworks: JAX, TensorFlow, PyTorch [Geometric], Triton, Keras, XGBoost, Scikit-Learn, AutoGluon

C/C++ Stack: CUDA, NCCL, MPI, OpenMP, libuv, Boost, gRPC, protobuf, MySQL, Hiredis, C++ REST SDK

Python Libs: FastAPI, Django, Flask, Gunicorn, SQLAlchemy, NumPy, Numba, Dask, Pandas, Matplotlib, Seaborn

Orchestration: Kubernetes, Ray, Slurm, Kubeflow, Airflow, Docker, TFX, TensorFlow Serving, Prometheus, Grafana

Cloud Platforms: Architecting with Google Kubernetes Engine, AWS Solution Architect Associate

Coursework: Language Modeling from Scratch (Stanford CS336), Deep Generative Models (Stanford CS236), Deep Learning (Berkeley CS182), Deep Reinforcement Learning (Berkeley CS285), Machine Learning with Graphs (Stanford CS224W), Intro to Computer Systems (CMU 15-213), Concurrent and Distributed Systems (U Cambridge)

Research Experience

Colorado School of Mines Golden, CO

Research Assistant *Jun 2014 – Aug 2019*

- Developed numerical simulation of patterning of non-spherical colloids under electric fields.
- Teamed with collaborators to design an $\mathbf{O}(N_b \log N_b)$ framework to model a large number of colloids.
- Simulated in- and out-of-equilibrium behaviors of particles in a high-performance computing cluster.