1775 Tysons Blvd, Ste 4172
Tysons, VA, 22102, USA

Shawn (Xiangru) Lian

shawn@shawnlian.com
https://shawnlian.com

Summary

Shawn is a highly experienced technology professional with a passion for engineering and algorithms. Prior to joining a stealth startup in 2022, He was a founding member of Kuaishou Technology (SEHK:1024) AI Platform, where he directed its Engineering with emphasis on Personalization Infrastructure and successfully led multiple core teams to drive innovative solutions. With his expertise, he developed the first GPU-based large scale advertising recommendation system at Kuaishou, which generated a revenue of 6 billion USD annually. In 2021, he created the world's largest scale recommender system (ACM SIGKDD 2022). He also served as a Senior Staff Research Scientist at Kuaishou Seattle AI Lab, where he was instrumental in inventing and developing cutting-edge systems utilizing the latest technologies, including large scale storage engines, high performance deep learning infrastructure, and model compression frameworks. His strong leadership and expertise in the field make him a valuable asset to any team.

Education

Ph.D. Computer Science **University of Rochester** **2015-2019**

- Advisor: Ji Liu
- GPA: 4.00/4.00, with exceptional performance (A+) in Advanced Algorithms, Operating Systems, Dynamic Languages and Software, Computational Complexity, Internship, and PhD Research

Visiting Student Researcher **Stanford University** **2014 Summer**

- Advisor: Zhi-Xun Shen
- Primary site: Geballe Laboratory for Advanced Materials (GLAM)

B.S. Applied Physics **University of Science and Technology of China** **2011-2015**

- Advisor: Tao Wu
- Yan Jici Talent Program in Physics
- GPA: 4.08/4.30. Ranking: 1/63 (Class 003) and 3/264 (School of the Gifted Young)

Employment

Stealth Startup (Washington D.C. Metropolitan Area)

- 2022.02 - : Principal Scientist & Vice President of Engineering

Full-Stack Technology Leadership & AI Innovation

- **Comprehensive Multi-Media Intelligence Platform**
  - Architected and led the development of the world's most comprehensive multi-media database encompassing books, movies, TV shows, and digital content with advanced AI-powered analysis capabilities. The platform processes petabytes of data with real-time inference and sophisticated content understanding across multiple modalities.
  - Designed and implemented the complete technology stack from ground up, demonstrating full-spectrum engineering expertise across frontend (React-based web applications), backend (high-performance Rust modular monolith), and cutting-edge AI systems for multi-modal content analysis.
- **Advanced AI & Machine Learning Systems**

  - Pioneered next-generation large language model (LLM) integration with proprietary image and video understanding algorithms, achieving state-of-the-art performance in content scoring and classification across diverse media types.
  - Developed revolutionary AI translation systems with context-aware multi-modal embedding techniques, enabling seamless cross-language content discovery and analysis with unprecedented accuracy.
  - Built distributed AI inference and training infrastructure supporting real-time processing of millions of media items with advanced computer vision and natural language processing pipelines.
- **Massive-Scale Data Engineering & Infrastructure**
  - Engineered world-class web scraping infrastructure capable of processing terabytes of data daily from thousands of sources with intelligent rate limiting, content deduplication, and automated quality assessment.
  - Designed fault-tolerant distributed storage systems with automated data lifecycle management, ensuring 99.99% uptime and seamless horizontal scaling across geographic regions.
  - Implemented sophisticated job scheduling and orchestration frameworks handling millions of concurrent tasks with dynamic resource allocation and intelligent workload balancing.
- **Cloud-Native & Security Architecture**
  - Architected innovative cross-cloud deployment strategy spanning multiple cloud providers and on-premise infrastructure, utilizing Kubernetes clusters for seamless workload portability and vendor independence.
  - Developed proprietary decentralized encrypted VPN mesh network with zero-trust security model, ensuring secure communication and data protection across distributed global infrastructure.
  - Implemented comprehensive DevOps practices with automated CI/CD pipelines, infrastructure as code, and advanced monitoring systems achieving deployment frequencies of 100+ releases per day.
- **Leadership & Team Excellence**
  - Successfully led cross-functional teams of 20+ engineers, designers, and researchers, fostering innovation culture while maintaining rapid development velocity and high-quality code standards across all technology domains.
  - Established technical vision and architectural standards spanning frontend user experience, backend scalability, AI model development, data engineering, and infrastructure operations, ensuring cohesive system design and optimal performance.

Kwai Inc. (Kuaishou Technology)

- 2019.08 - 2022.01: Founding Member of AI Platform & Senior Staff Research Scientist
- 2018.10 - 2018.12: Research Intern

Selected Projects

- **PERSIA**
  - Kuaishou's advanced GPU-based large scale learning system designed for ad recommendation and CTR prediction tasks. Launched in 2018 by me, PERSIA has been leading the charge in the field of recommendation systems, and was open-sourced in 2021. With the ability to support models with up to 100 trillion parameters, PERSIA is the fastest public recommendation model training framework available. Built using Rust for high performance computing and communication, PERSIA is a testament to Kuaishou's commitment to advancing the state of the art in recommendation systems.
  - The system supports commercialization department to generate 6 billion USD annually and the oversea department to grow more than 100 million daily active users.
  - Received the Tech Breakthrough Award from Kuaishou Technology. Broadly covered by media including Synced, AI Front, China Daily, and InfoQ.
- **Bagua**
  - Kuaishou's deep learning training acceleration framework designed to tackle the challenge of large scale training tasks. Bagua offers a comprehensive solution to speed up the training process, including data loader optimization, advanced distributed training algorithms, network communication optimization, and more. Developed to solve the training bottleneck at Kuaishou Technology, where more than a million videos are uploaded every hour, Bagua has been instrumental in maintaining the company's position at the forefront of innovation.
  - Bagua can be faster than existing solutions like Horovod and BytePS by more than 50%.

  - Bagua's introduction is selected on NVIDIA GTC as one of five Simulive Sessions, and broadly covered by media (I list a few here): AI Front, InfoQ, and Rust Chinese Community.
  - Bagua is now open sourced on GitHub.
- **Hammer**
  - The automatic deep learning model compression tool developed by Kuaishou Technology. With Hammer, reducing the size of large models while maintaining their accuracy has never been easier. This innovative tool has already made a big impact at Kuaishou, saving thousands of GPU cards and enabling the successful deployment of hundreds of complex models. (Hammer also helped the TAMU-KWAI team won the 2nd prize in IEEE Low-Power Computer Vision Challenge).
  - Media coverage: NVIDIA GTC, and Kuaishou Technology.
- **DouZero**
  - DouZero is our game AI for DouDizhu. A small scale version has been released on GitHub and is covered by multiple media like Synced. The corresponding research paper is accepted by ICML 2021. My team helped build the DouZero GPU based distributed reinforcement learning training platform, to support large scale training of game AI.

IBM Thomas J. Watson Research Center

- 2018.05 - 2018.08: Graduate Research Internship
- 2017.05 - 2017.08: Research Summer Intern

Selected Projects

- **Large scale decentralized asynchronous parallel training for deep learning**
  - The DSPGD/ADPSGD algorithms are revolutionary decentralized training solutions for artificial intelligence, offering unparalleled speed and accuracy. Unlike traditional training algorithms, these decentralized algorithms are optimized for cloud computing environments where network conditions and machine performance can vary. As demonstrated by IBM, the DSPGD/ADPSGD algorithms can drastically reduce training times for speech recognition AI, from a week to just 11 hours, while also delivering a 10x improvement in performance.

Tencent AI Lab (Tencent America LLC)

- 2017.09 - 2017.12: AI Research Intern

Selected Projects

- **Game AI for King of Glory**
  (The King of Glory is the most popular MOBA game in China, with 160 million daily active users.)
  - Design and implement an efficient learning platform called KOG-Sgame which supports training both supervised learning based and reinforcement learning based game AI. The platform is later used in the whole Seattle Game AI team to develop game AI algorithms.
  - The first AI defeating the internal rule based AI is trained by me on this platform.
  - Co-design the game arena for different King of Glory game AIs to compete with each other.
  - Before leaving, for 1 vs 1 games our AI matches top 10% human performance.

Awards & Honors

- **2019**
  - 30 New Generation Digital Economy Talents
  - Kwai Inc. 2019 Q2 Technology Breakthrough Award
- **2018**
  - Intel Student Ambassador
  - ICML Travel Award
- **2017**
  - NIPS Travel Award
- **2015**

  - NIPS Travel Award
  - Outstanding Undergraduate of University of Science and Technology of China
  - Honorary Rank of Academic Achievement of the Grade 2011 Undergraduates
  - Honor of Graduating from Yan Jici Talent Program in Physics
- **2014**
  - National Scholarship
- **2013**
  - National Scholarship
  - 1st Prize of 5th China Undergraduate Mathematical Contest (Anhui division)
  - Grand Prize of USTC Research Oriented Physics Experiment Competition
  - 2nd Prize of USTC Undergraduate Mathematical Contest
- **2012**
  - National Encouragement Scholarship
  - 1st Prize of 4th China Undergraduate Mathematical Contest (Anhui division)
  - 2nd Price of "USTC Star" Forum LOGO Design Competition
- **2011**
  - USTC Scholarship for Outstanding Freshman

Publications

(* means equal contribution)

Refereed Conference Proceedings & Journal Articles

Xiangru Lian, Binhang Yuan, Xuefeng Zhu, Yulong Wang, Yongjun He, Honghuan Wu, Lei Sun, Haodong Lyu, Chengjun Liu, Xing Dong, et al. "Persia: A Hybrid System Scaling Deep Learning Based Recommenders up to 100 Trillion Parameters". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2022

Shaoduo Gan*, **Xiangru Lian***, Rui Wang, Jianbin Chang, Chengjun Liu, Hongmei Shi, Shengzhuo Zhang, Xianghong Li, Tengxu Sun, Jiawei Jiang, Binhang Yuan, Sen Yang, Ji Liu, and Ce Zhang. "BAGUA: Scaling up Distributed Learning with System Relaxations". In: *VLDB*. 2022

Xiao Hu, Ming-Ching Chang, Yuwei Chen, Rahul Sridhar, Zhenyu Hu, Yunhe Xue, Zhenyu Wu, Pengcheng Pi, Jiayi Shen, Jianchao Tan, **Xiangru Lian**, Ji Liu, Zhangyang Wang, Chia-Hsiang Liu, Yu-Shin Han, Yuan-Yao Sung, Yi Lee, Kai-Chiang Wu, Wei-Xiang Guo, Rick Lee, Shengwen Liang, Zerun Wang, Guiguang Ding, Gang Zhang, Teng Xi, Yubei Chen, Han Cai, Ligeng Zhu, Zhekai Zhang, Song Han, Seonghwan Jeong, YoungMin Kwon, Tianzhe Wang, and Jeffery Pan. "The 2020 Low-Power Computer Vision Challenge". In: *3rd IEEE International Conference on Artificial Intelligence Circuits and Systems, AICAS 2021*

Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, **Xiangru Lian**, Xia Hu, and Ji Liu. "DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning". In: *International Conference on Machine Learning*. 2021

Hanlin Tang, Shaoduo Gan, Ammar Ahmad Awan, Samyam Rajbhandari, Conglong Li, **Xiangru Lian**, Ji Liu, Ce Zhang, and Yuxiong He. "1-bit Adam: Communication Efficient Large-Scale Training with Adam's Convergence Speed". In: *International Conference on Machine Learning*. 2021

Wenqing Hu, Chris Junchi Li, **Xiangru Lian**, Ji Liu, and Huizhuo Yuan. "Efficient Smooth Non-Convex Stochastic Compositional Optimization via Stochastic Recursive Gradient Descent". In: *Advances in Neural Information Processing Systems*. 2019

Hanlin Tang, **Xiangru Lian**, Tong Zhang, and Ji Liu. "`DoubleSqueeze`: Parallel Stochastic Gradient Descent with Double-Pass Error-Compensated Compression". In: *International Conference on Machine Learning*. 2019

Xiangru Lian and Ji Liu. "Revisit Batch Normalization: New Understanding and Refinement via Composition Optimization". In: *Proceedings of Machine Learning Research*. Vol. 89. Proceedings of Machine Learning

Research. PMLR, 2019, pp. 3254–3263

Hanlin Tang, **Xiangru Lian**, Ming Yan, Ce Zhang, and Ji Liu. "D^2: Decentralized Training over Decentralized Data". In: *International Conference on Machine Learning*. 2018

Xiangru Lian*, Wei Zhang*, Ce Zhang, and Ji Liu. "Asynchronous Decentralized Parallel Stochastic Gradient Descent". In: *International Conference on Machine Learning*. 2018 **(long talk)**

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. "Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent". In: *Advances in Neural Information Processing Systems*. 2017 **(oral paper, rate 1%)**

Xiangru Lian, Mengdi Wang, and Ji Liu. "Finite-sum Composition Optimization via Variance Reduced Gradient Descent". In: *International Conference on Artificial Intelligence and Statistics*. 2017

Yang You*, **Xiangru Lian***, Ji Liu, Hsiang-Fu Yu, Inderjit Dhillon, James Demmel, and Cho-Jui Hsieh. "Asynchronous Parallel Greedy Coordinate Descent". In: *Advances in Neural Information Processing Systems*. 2016

Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. "A Comprehensive Linear Speedup Analysis for Asynchronous Stochastic Parallel Optimization from Zeroth-Order to First-Order". In: *Advances in Neural Information Processing Systems*. 2016

Wei Zhang, Suyog Gupta, **Xiangru Lian**, and Ji Liu. "Staleness-aware Async-SGD for Distributed Deep Learning". In: *International Joint Conference on Artificial Intelligence*. 2016

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. "Asynchronous parallel stochastic gradient for nonconvex optimization". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2719–2727 **(spotlight paper, rate 4%)**

Yongping Wu, Dan Zhao, **Xiangru Lian**, Xiufang Lu, Naizhou Wang, Xigang Luo, Xianhui Chen, and Tao Wu. "NMR evidence for field-induced ferromagnetism in $(Li_{0.8}Fe_{0.2})$OHFeSe superconductor". In: *Physical Review B* 91.12 (2015), p. 125107

Preprints

Huizhuo Yuan, **Xiangru Lian**, and Ji Liu. *Stochastic Recursive Variance Reduction for Efficient Smooth Non-Convex Compositional Optimization*. 2020. arXiv: `1912.13515 [stat.ML]`

Hanlin Tang, Shaoduo Gan, Samyam Rajbhandari, **Xiangru Lian**, Ji Liu, Yuxiong He, and Ce Zhang. *APMSqueeze: A Communication Efficient Adam-Preconditioned Momentum SGD Algorithm*. 2020. arXiv: `2008.11343 [cs.DC]`

Huizhuo Yuan, **Xiangru Lian**, Ji Liu, and Yuren Zhou. *Stochastic Recursive Momentum for Policy Gradient Methods*. 2020. arXiv: `2003.04302 [stat.ML]`

Professional Services

Reviewer/Program Committee

- **Senior PC**
  - AAAI Conference on Artificial Intelligence (2019)
- **PC/Reviewer**
  - International Conference on Machine Learning (ICML 2019, 2020, 2021)
  - Neural Information Processing Systems (NeurIPS 2016, 2019, 2022)
  - International Conference on Artificial Intelligence and Statistics (AISTATS 2022)
  - International Conference on Learning Representations (ICLR 2021, 2022, 2023, 2024)
  - AAAI Conference on Artificial Intelligence (AAAI 2018, 2020)
  - NeurIPS Reproducibility Challenge (2019)
  - Scalable Deep Learning over Parallel And Distributed Infrastructures (ScaDL 2019, 2020, 2021)

- Journal of Machine Learning Research
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- IEEE Transactions on Information Theory
- IEEE Transactions on Neural Networks and Learning Systems
- IEEE Transactions on Network Science and Engineering
- IEEE Transactions on Knowledge and Data Engineering
- IEEE Transactions on Signal Processing
- IEEE Internet of Things Journal
- Machine Learning
- International Journal of Electrical Power and Energy Systems
- Data Mining and Knowledge Discovery
- Measurement
- Neurocomputing
- Parallel Computing
- Neural Networks
- Pattern Recognition
- Optimization Methods and Software
- Computational Optimization and Applications
- Journal of Parallel and Distributed Computing
- European Journal of Operational Research
- BIT Numerical Mathematics
- Journal of Optimization Theory and Applications

Invited Talks

- **2021**:
  - RustMagazine (Chinese): Rust FFI: Distributed Deep Learning
  - DataFun Summit (Chinese): What is the Training Acceleration Framework that Breaks PyTorch/TensorFlow's Scalability Bottleneck?
- **2020**:
  - NVIDIA GPU Technology Conference: *Bagua! Distributed Communication Library*
- **2019**:
  - Kwai Inc.: *Technology best practices - Large scale GPU based learning system for ad recommendation*
  - ICML 2019: *DoubleSqueeze: parallel stochastic gradient descent with double-pass error-compensated compression*
  - Kwai Inc.: *On model compression and distributed training of the Face Landmarks model*
- **2018**:
  - INFORMS 2018 International Meeting: *Asynchronous Parallel Empirical Variance Guided Algorithms for the Thresholding Bandit Problem*
  - ICML 2018: *Asynchronous Decentralized Parallel Stochastic Gradient Descent*
  - ICML 2018: *D2: Decentralized Training over Decentralized Data*
- **2017**:
  - IBM T.J. Watson Research Center, Optimization for AI: *Accelerating Deep Learning via Decentralized Parallel Optimization*
- **2015**:
  - NIPS 3min spotlight talk: *Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization*

Teaching Assistant

- **2018**: CSC446 – Machine Learning
- **2017**: CSC484 – Advanced Algorithms
- **2016**: CSC282 – Design and Analysis of Efficient Algorithms

Technologies

GNU/Linux is my primary operating system for nearly two decades. I like to self-host services I use (some written by myself) on my own Kubernetes cluster. I have a strong background in polyglot programming and a proven track record of quickly adapting to new technologies. I understand a diverse range of over 20 programming languages such as:

> Rust, ReasonML (OCaml), F#, Clojure, ClojureScript, Java, TypeScript, Hy, Python, C, Racket, Emacs Lisp, Groovy, Haskell, SQL, AQL, LogQL, GraphQL, Julia, Solidity, Matlab, Mathematica, C++, C#, Go, Lua, Ruby, LaTeX, ConTeXt, Bash, Zsh, Fish, LabVIEW, HTML, CSS, Sass, and PHP.

With a hands-on approach to learning, I have developed my skills through practical experience and managing teams of varying technical backgrounds. This expertise allows me to quickly pick up any new technology and put it into practice.

I am familiar with but not limited to the following list of technologies.

- **Deep learning**
  - PyTorch, TensorFlow, and Swift for TensorFlow (sadly this project is abandoned now)
  - Creating custom higher performance operator with tensor libraries like ArrayFire or General-Purpose GPU programming such as CUDA
  - MPI, NCCL and more for distributed communication
- **Microservices/distributed systems**
  - **Communication**
    - Existing higher level tools like gRPC, ZeroMQ
    - Underlying technologies like TCP and RDMA
    - Created new RPC systems for extremely performance-demanding scenarios
  - **Service mesh/discovery**: Consul, Linkerd, ...
  - **KV**: Redis, Consul KV, etcd, ZooKeeper, ...
  - **Secret management**: Vault
  - **Orchestration** Kubernetes, Nomad, MRSK, ...
  - **Virtualization/CGroup/AUFS** KVM, Docker, podman, LXD, LXC, systemd-nspawn, Firecracker, ...
  - **Observability** Sentry, Grafana, Prometheus, Tempo, OpenTelemetry, ...
  - **"BigData"** HDFS, YARN, Hive, Spark, ... (But you don't need big data, you need the right data :)
- **Database** PosgreSQL, CockroachDB, Sqlite, EdgeDB, ScyllaDB, ...
- **High performance computing (mainly on x86-64)** SIMD vectorization, non-blocking I/O (with `io_uring` or `epoll`), lockless data structures, allocators, cooperative multitasking, GPGPU programming, ...
- **Embedded programming** STM32 ARM programming with Rust using `no_std`
- **DevOps** CI/CD (GitLab pipeline w. custom runner, GitHub Actions, ArgoCD, Buildkite, DroneCI, ...), IaC (Terraform, Pulumi, Ansible)
- **Web development**
  - **Frontend** React/ReasonReact, TanStack, Astro, Remix, Svelte, Yew (compiled to WebAssembly), re-frame, jotai, Next.js
  - **Backend** Too many, every languages has its own HTTP service implementation
  - **Functional** Effect (TypeScript)
  - **Interface** GraphQL, REST
- **Desktop development** GTK, and Tauri
- **Mobile development**
  - Android native application development (Kotlin, Jetpack Compose, MVVM/MVI, JNI, ...)
  - Hooking with Xposed and its derivatives
  - Cross platform mobile application development with Flutter / ClojureScript (ReactNative)
- **UI/UX** Figma
- **Reverse engineering** Static analysis, dynamic analysis, binary instrumentation, and protocol analysis