Contact

Sunnyvale, California, USA
zyh828482@gmail.com

www.linkedin.com/in/yaohai-zhou (LinkedIn)
www.yaohaizhou.com/ (Blog)
github.com/zyhhhy (Personal)

Top Skills

Qdrant
Golang
Gin

Languages

English (Full Professional)
Chinese (Native or Bilingual)

Publications

Channel Prediction with Liquid Time-Constant Networks: An Online and Adaptive Approach

Efficient Human Pose Estimation by Maximizing Fusion and High-Level Spatial Attention

John Zhou

Seeking 24NG Entry SDE | MEng (CS/AI) @ UCLA | Sunnyvale | Intel
San Francisco Bay Area

Summary

I am a CS MS candidate at UCLA with much programming experience, especially in full-stack development, CloudOps and machine learning. I have solid programming skills in Golang, Python, C++, Java, and JavaScript. And I have hands-on experience with Gin, React, Vue, Flask, PyTorch, Docker, Kubernetes, Terraform, Kafka, GitHub, AWS, GCP, and other cloud technologies. I have a passion for innovation, a thirst for technologies, and a strong sense of collaboration.

Experience

PPIO
Software Engineer Intern
February 2023 - July 2023 (6 months)
Los Angeles, California, United States

DazzleAI (https://www.dazzleai.ai/): Top-10 Stable-Diffusion (SD) Text-to-Image Generation Web Platform with 10k Users in 2 Months

• Web Development. Implemented web frontend using React, backend using Go and Python in MicroService architecture, including task distributed system using AWS SQS message queue, image sharing community, Google login, Stripe subscription, daily checkin. Used AWS and GitHub Actions to perform CI/CD pipelines, GitHub Webhook to notify the code change events.
• Deployment. Deployed the website online using AWS Elastic Beanstalk, IAM, S3, MySQL, SQS, SES, SNS, Cloudfare CDN.
• Virtualization. Utilized Docker to streamline environment setup and used InfluxDB and Grafana for system operating monitor.
• Container Orchestration. Orchestrated multiple Stable-Diffusion Docker containers using Kubernetes for efficient scaling.
• Impact. Provided over 1200 Stable-Diffusion models on distributed GPUs and achieved text-to-image generation in 5 seconds.

Shenzhen innoX
Software Engineer Intern

February 2022 - April 2022 (3 months)
Shenzhen, Guangdong, China

• Used 2D camera pictures to generate vision maps using Multiple Vision Geometry and Computer Vision techniques.
• Collect raw images through fisheye cameras, preprocess and extract semantic segmentation result in Python and MM-Detection.
• Proposed a method to filter straight lines (poles and lights) in maps and corresponding positioning algorithm in tensorRT C++.
• Deployed the model on a sweeper car and joint debugged with navigation and object detection algorithms.

Intel Corporation

Software Engineer Member | Project Core Member

July 2021 - January 2022 (7 months)
Beijing, China

AI Channel Link Adaptation: Enhancing Transmission Throughput in 5G Scenarios using Reinforcement Learning (RL)
• Experiment Environment Setup. Established an offline data pipeline and designed the training and testing environment with PyTorch. Utilized Matlab for simulating ideal wireless transmission scenarios and employed Python for data visualization.
• RL Algorithm Research & Development.
⋆ Innovatively integrated Dueling Double DQN and Time Series Neural Networks to handle multi-frame data effectively.
⋆ Engineered an efficient input combination to expedite model adaptation to 5G wireless channel changes.
⋆ Applied mathematical techniques to accelerate convergence speed, aligning with link adaptation principles.
• Model Deployment. Converted model to ONNX and deployed the model on Intel’s FlexRAN platform using OneDNN in C/C++.
• Impact. Achieved a 10% increase in transmission throughput and faster adjustment speed than traditional algorithms.

Huazhong University of Science and Technology Dian Group

Software Engineer Intern

May 2020 - June 2021 (1 year 2 months)
Wuhan, Hubei, China

Truck Detection: Build Real-Time Truck Detection and Monitor System in Camera Image Streaming Processing Pipeline

• Model Enhancement. Optimized object detection model by model distillation, model fusion, attention mechanism in Python.
• Streaming Pipeline. Implemented Kafka and Flink in Java to build a seamless online image streaming processing pipeline.
• Web Development. Created 20+ RESTful APIs using Vue and Flask to build functionality between the frontend and backend.
• Impact. Achieved 96% accuracy and 93% recall on real truck detection scenario. Saved 50% human labor to detect trucks.

Education

University of California, Los Angeles

Master of Engineering - MEng, Artificial Intelligence/Computer Science · (September 2022 - December 2023)

Huazhong University of Science and Technology

Bachelor of Engineering - BE, Electrical and Computer Engineering · (September 2018 - June 2022)