Xingchen Zhao

Results-oriented full-stack engineer, looking for Software Engineer Position

+1-571-525-4502 tinyurl.com/bddsvxzh zhao.xingc@northeastern.edu linkedin.com/in/xingchenzhao/

Education

Northeastern University — **Boston, MA**
Master of Science in Computer Engineering, GPA: 3.67/4.0 — August 2021 - April 2024
PhD in Computer Engineering — (Dropout) August 2021 - January 2023

University of Pittsburgh — **Pittsburgh, PA**
Bachelor of Science in Computer Science, GPA:3.71/4.0 — January 2018 – May 2021

UC San Diego, ML Visiting Scholar — November 2020 - May 2021

Working Experience

Software Engineer Intern — **August 2023 – Jan 2024**
Amazon Robotics — Westborough, MA

- Led development of light-weight real-time object detector for AR-ID Localizer, enhancing performance by **5%** mAP and increasing speed by **120%** on edge devices, achieved by distribute training on 40k shipping label images.
- Designed unified model for localizing 1D/2D barcodes on Amazon packages and vendor products, improving accuracy by 3%. Facilitated efficient package sorting, enhancing delivery effectiveness for billions of customers.
- Deployed models in **ONNX** format on edge device, convert to **TensorRT** for optimized FP16 and Int8 inferencing.
- Used SageMaker CI/CD pipelines to build, integrate, test, and distribute optimized models, reduced 40% workload.

Founding Software Engineer — **February 2023 – Now**
Learnie AI | develop.learnieai.com | Partnered with the US's largest edu company, HMH — Boston, MA

- Invented and developed full-stack K-12 edu platform, using React and Node.js with TypeScript, and PyTorch for Deep Learning, and Firebase for database and storage. Enhanced personalized learning experience for 500+ students.
- Engineerd talking-head API, boosted speed by 65%, utilized Docker, Redis caching, and Grafana monitoring.
- Architected the backend with NLP, Speech Recognition, Text-to-Speech using ChatGPT, Azure TTS, and fine-tuned **LLM** model; set up automated testing via PyTest, orchestrated services via Kubernetes, and deployed via Vercel.
- Led an agile team of 6 in sprint planning and daily sync stand-ups, utilizing Jira and GitHub.

Founding Software Engineer — **April 2023 – Now**
Chat AI Zoo | chataizoo.com — Boston, MA

- Led team of 10, scaling the AIGC social platform for user-generated LLM agents to 1K daily users and 100K clicks.
- Developed SpringBoot/Flask services, containerized by Docker and managed by Kubernetes with 99.9% uptime.
- Reduced data retrieval latency by 70% with a real-time chat history pipeline using Redis, DynamoDB, and Pinecone;
- Orchestrated AWS infrastructure via Terraform, including EC2 and ECR, reducing setup time by 30%.
- Optimized Nginx with Elastic Load Balacing for distributed traffic and high availability, reduced 30% response time.

Machine Learning Engineer Intern — **May 2022 – August 2022**
SRI International — Princeton, NJ

- Developed label-free segment model, saved 50% on costs; won 1st in DarkZurich with self-supervised algorithm.
- Used **DeepSpeed** and ZeRO for distribute training with model/data parallelism, achieving a **7%** accuracy boost.
- Enhanced model efficiency by **90%** using Knowledge Distillation; deployed on AWS using ONNX and TensorRT.

Machine Learning Researcher — **January 2020 – January 2023**
Northeastern University, University of Pittsburgh, UCSD — Boston, Pittsburgh, San Diego

- Authored 8 ML papers on CV, NLP, and Multimodality; reviewed 20+ papers in CV, NLP domains.
- Developed **multi-modal** techniques with Transformers, boosting object detection accuracy by 4.5% over SOTA.

Selected Publications | Machine Learning/Artificial Intelligence | 7 Paper Published

- **Zhao, X.**, Sicilia, A., ..., "Test-time Fourier Style Calibration for Domain Generalization", IJCAI, 2022.
- **Zhao, X.**, Minhas, D., ..., "Robust White Matter Hyperintensity Segmentation on Unseen Domain", ISBI, 2021.

Technical Skills

Machine Learning: Neural Language Processing, Computer Vision, Multimodal, Object Detection, Segmentation
Programming languages: Python, C++, Java, C, R, JavaScript, HTML, CSS, Swift, SQL
Software Frameworks: PyTorch, TensorFlow, DeepSpeed, TensorRT, Huggingface, Spark, OpenCV, React, NodeJS
DevOps and Cloud Technologies: Docker, Git, Kubernetes, Google Cloud, AWS, Apache Spark