Botong Ou

bou@purdue.edu | 765-586-0756 | Personal Website | Github | Linkedin

EDUCATION

Purdue University - Main Campus (Ph.D. quit) May. 2024 (Expected)
Master's degree in Computer Information Technology

University of California, Los Angeles (UCLA) Sept. 2019 - May. 2021
Master's degree in Computer Science

Shanghai Jiao Tong University (SJTU) Sept. 2015 - May. 2019
Bachelor's degree in Computer Science

SKILLS

Programming Languages	Java, Python, C/C++, Golang, JavaScript, Rust, HTML/CCS
Frameworks	ReactJS, Django, Nginx, Redis, Docker, Kubernetes, MongoDB, Springboot
Features	Full Stack Development, LLM, MLOps, AIGC, Database

WORK EXPERIENCE

LetsRent LLC - Los Angeles May. 2023 - Present
Software Engineer Intern

- Engineered a robust service platform that seamlessly integrates large language models with expansive SQL databases.
- Major Contributions:
  * Adopted **Django** to build web server for hosting bespoke **LLM** models such as 8-bit quantized Vicuna model.
  * Employed **ReactJS** in constructing responsive web pages that displayed events via **RESTful API** requests.
  * Utilized **Langchain** with LLM models for customizing a knowledge-based chat agent querying SQL database.
  * Improved database search speed by integrating **Pinecone** and **Chroma** vector database for cloud environment.
  * Leveraged **AWS Lambda** and Elastic Load Balancer for reliable, auto-scaling services during high traffic peaks.
- The system provides context-aware information by instructing large language models with precision-targeted prompts.

Tensorchord - Remote Dec. 2022 - Mar. 2023
Software Engineer Intern

- Developed container-based **MLOps** - **Envd** with integrated support of multiple languages and ML frameworks.
- Main Contributions:
  * Designed CLIs for users to provision ML environments in **Python/Julia/R** without manually adding dependencies.
  * Adopted remote and local caching to accelerate the build time by **4x** faster for customized ML environments.
  * Integrated with **Kubernetes** for distributing ML workloads with autonomous network configurations.
  * Introduced continuous integration and delivery (CI/CD) to facilitate the deployment using **Github Actions**.
  * Incorporated **Logstash** for real-time monitoring and logging, improving observability of ML workflows in Envd.
- Envd has received **>1700** stars in MLOps community and obtained **>5000** users by the end of 2022.

RSSys - Purdue University Sept. 2021 - May. 2022
Research Assistant

- Proposed the state-of-art Confidential Virtual Machine (**CVM**) architecture against untrusted cloud infrastructure.
- Major Contributions:
  * Designed **Slab** memory allocation algorithm for **Library OS** to reduce memory fragmentation.
  * Developed an audit log system monitor to store **~1G** system logs information in a reserved memory region.
  * Supported various runtime for applications including **Redis**, **Nginx** and **OpenSSL** with **10% - 15%** overhead.
- The work is accepted at the incoming **ASPLOS 2024** - CCF Class A conference in the computing infrastructure field.

NESL - University of California, Los Angeles Sept. 2019 - May. 2021
Research Assistant

- Designed the first edge system that provides fast deep learning inference for mobile and IoT devices.
- Major Contributions:
  * Deployed **MongoDB** database on edge device to collect data generated locally at the speed of 20G daily.
  * Leveraged Google's **OpenThread** network protocol to allow **AD-HOC** communication between cloud containers.
  * Allows **>500** containers to transmit data between each other with only **~80ms** latency introduced.
  * Supported multiple modern ML/DL models to run on the edge devices with **~5%** performance overhead.
- The work is accepted by **IoTDI 2021** - CCF Class A conference in the IoT field and has received **>200** citations.