YUCHEN ZHENG

zhengyuchen1017@ucla.edu | Website: alexzheng.io | Linkedin: alexzheng1017 | GitHub: Alex Zhang-UCLA

EDUCATION

UNIVERSITY OF CALIFORNIA, LOS ANGELES (UCLA) *Expected Mar 2024*
MS in Electrical & Computer Engineering, GPA: 3.90/4.0
Courses: Natural Language Processing, Matrix Analysis, Large-scale Data Mining, Large-scale Network

ZHEJIANG UNIVERSITY *Sep 2018 - Jun 2022*
BE in Robotics, Minor in Management and Entrepreneurship, GPA: 3.88/4.0, top 5%
Courses: Programming, Data Structure, Robotics, Machine Learning

SKILLS

Programming Languages	Python, C++, Java, JavaScript, HTML/CSS, Linux command
Frameworks and Libraries	Node.js, Express, PyTorch, TensorFlow, Flask, Fastapi, Redis, React, Gradio, Streamlit
Databases	MySQL, MongoDB, Google Firebase
Tools and Software	VSCode, PyCharm, Git, HuggingFace, Docker, AWS EC2 & S3, Colab, Slack

EXPERIENCE

SOFTWARE ENGINEER INTERN *July 2023 - Sep 2023*
HUAWEI | PYTHON, FASTAPI, DOCKER, MONGODB, CLOUD PLATFORM, GIT, TRANSFORMERS

- Architected inference pipeline of Llama-like LLMs and embedding models as api services leveraging **FastAPI**, **Docker, MongoDB.** Deployed on internal **cloud platform** to bolster data security and preservation
- Integrated retrieval-augmented LLM enpowered by **LangChain** and vector database seamlessly into office software through API calling, leading to a 'Team Docs Helper' tool which managed 100+ daily queries in beta
- Refactored over 200 **Python** business code files to meet Huawei 'clean code' principles; collaborated via **version control,** ensuring a timely and error-free codebase delivery to clients

SOFTWARE ENGINEER INTERN *Jan 2023 - June 2023*
VITALLY (AI STARTUP) | PYTHON, FLASK, REDIS, GRADIO, REDIS QUEUE, CLOUD PLATFORM, GIT, HUGGING FACE

- Optimized machine learning code framework integrated with **Gradio** UI interface, facilitating machine learning engineers in crafting custom image-generation models
- Implemented **Flask** backend paired with **Redis Queue** workers for cloud-based distributed task management, enhancing AIGC application performance and reducing GPU resource costs by over 50%
- Led the development of an object detection module within a broader application pipeline, leveraging top models from **HuggingFace** to attain an accuracy of over 95%

MACHINE LEARNING ENGINEER INTERN *July 2021 - July 2022*
RESEARCH INSTITUTE OF ZHEJIANG UNIVERSITY | C++, PYTHON, PYTORCH, UBUNTU, ROS, PANDAS, LATEX

- Spearheaded large-scale driving data collection and noisy datasets preprocessing; designed and trained custom deep learning model using **PyTorch**, achieving a modeling error rate of under 10%
- Seamlessly integrated neural network-driven vehicle models into the complete **C++** autonomous driving system on **Ubuntu**, culminating in a 16% reduction in control error
- Awarded the 'Outstanding Graduate' honor for graduation thesis derived from segments of this work

PROJECTS (END-TO-END DEVELOPMENT)

AIGC Experience App , Full-stack App | PYTHON, STREAMLIT, FASTAPI, REDIS, AMAZON EC2 & S3 *June 2023 - Now*
Machine learning app that let users easily experience hot AI models including text/image generation, audio2txt

Stable Diffusion Model Trainer | PYTHON, PYTORCH, GRADIO, GIT *April 2023 - June 2023*
Stable diffusion webui extension that solve dependencies issue and easy to use for stable diffusion model training