Jimmy Shong

(650) 681-7291 • jimmysh341@gmail.com • https://jiminator.github.io/PersonalSite/

EDUCATION

University of Illinois Urbana-Champaign, Urbana, IL Aug 2024-Present

M.S. in Computer Science

New York University Tandon School of Engineering, Brooklyn, NY Sep 2020-May 2024

B.S. in Computer Science, Minors in Math and Cybersecurity, GPA: 3.8/4.0, Dean's List 2020-2023

RESEARCH INTERESTS

Investigations into improving speed, efficiency, and performance of machine learning models and systems through model compression, efficient model serving, and inference optimization. I am also interested in privacy-preserving ML and secure deployment of ML models.

RESEARCH EXPERIENCE

Massachusetts Institute of Technology, MIT HAN Lab, Cambridge, MA May 2023-Present

Research Intern

- Implemented techniques such as loop unrolling, multi-threading, and SIMD programming to optimize matrix multiplication for TinyChatEngine, a high-performance LLM inference library.
- Built **TinyVoiceChat**, a local chatbot that utilizes TinyChatEngine, and Whisper to allow users to interface with a quantized LLM on edge devices solely with their voice.
- Researched matrix multiplication backends targeting Apple Silicon devices, specifically the Apple Neural Engine and Apple GPU through the CoreML and Metal frameworks, respectively.

NYU AI for Scientific Research, Brooklyn, NY Sep 2021-May 2022

Unsupervised ML Lead

- Engineered a Python library employed by researchers at the University of Groningen that utilizes supervised ML algorithms to analyze macrophage trajectories and predict their diffusion state.
- Performed data visualization and analysis on the sleep-wake dynamics of mice for scientists at NYU Abu Dhabi.

PROJECTS

Exploring SFT Methods LLMs (PyTorch) May 2024

- Instruct-tuned Llama3-8B on SlimOrca and Nectar datasets using supervised finetuning (SFT) methods such as LoRA, QLoRA, LoRA+, and Badam with Llamafactory.
- Evaluated the training speed, memory usage, and model performance of these SFT methods.

Google Suite Task Manager (Flask, MongoDB) May 2024

- Developed a task manager that aggregates all tasks created in both Google Tasks and other Google Suite products.
- Implemented task management features currently available in Google Tasks, including creation and deletion of tasks and deadlines.
- Connected the application to a MongoDB database stores all tasks and relevant task metadata

Efficient ResNet (PyTorch) March 2024

- Created a <5M modified ResNet architecture that achieves a 96.1% accuracy on CIFAR-10 dataset, which is a 3% increase in accuracy over the reported 11.4M ResNet-18 model.
- Performed ablation studies on different neural network optimizations including pooling, dropout, data augmentations, schedulers, and optimizers.

PROJECTS CONTINUED

NetArmor (Python, TurboGears, SQLAlchemy) August 2023

- Designed web-based tool that allows website owners to run scans on their websites, identify vulnerabilities, and connect with certified cybersecurity professionals to patch those issues.
- Engineered a PostgreSQL database using SQLAlchemy to store all necessary data and implement endpoints that the frontend uses to communicate with the database.

Air Ticket Reservation System (Flask, HTML, MySQL) December 2022

- Developed a web-based application that allows airline customers and staff to perform all necessary functions to operate an airport flight transaction system.
- Built a relational database built with to store necessary data and handle all transactions.

Face Recognition Attendance System (Python, OpenCV, DLib) June 2021

- Created a tool that uses computer vision libraries to take attendance of a class using a video feed.

TEACHING EXPERIENCE

NYU Polytechnic Tutoring Center, Brooklyn, NY Jan 2023-May 2024

Computer Science Tutor

- Helped NYU CS students with questions regarding required computer science courses.
- Recorded explanation videos for answer keys of mock exams developed by the PTC.

BlueStamp Engineering, Palo Alto, CA Jun 2022-Jul 2022

Instructor

- Taught practices and principles of engineering to a class of twenty high school students.
- Led projects such as an intelligent door lock, a smart mirror, and Alexa home automation.

TECHNICAL SKILLS

- Programming Languages: Python, C++, C, Java, JavaScript, HTML, CSS, MySQL
- Frameworks and Tools: PyTorch, Tensorflow, TurboGears, CoreML, Metal, Flask, TurboGears, SQLAlchemy, MongoDB, Git, AWS
- Other: Linux(Ubuntu, Kali), MacOS

LANGUAGES

English: Fluent
Mandarin: Intermediate