HAORUO ZHANG

Mobile：86-18691568778 Email：eric-zhang0628@hotmail.com

EDUCATION

University of California, San Diego 09/2022 – 06/2024
Master of Science CA, USA

- **Weighted Average Score:** 3.461/4.0
- **Major:** Computer Science & Engineering
- **Selected Courses (A/A+):** ML: Learning Algorithms, Unsupervised Learning, Neural Networks & Pattern Recognition, Statistical NLP, LLM AI & HCI, Advanced Computer Vision, Intro to Embedded Computing, Intro to Robotics, Application of Specific Processors, Parallel Computer Architecture

University of California, Berkeley 09/2018 - 05/2022
Bachelor of Science CA, USA

- **Major:** Computer Science
- **Weighted Average Score:** 3.621/4.0
- Dean's Honor List – College of Letters & Science (Top 10%) Spring 2021
- **Selected Courses (A/A+):** Principles & Techniques of Data Science, Issues in Cognition, Intro to CS Theory, Artificial Intelligence, Database Systems, Data Structures

RESEARCH INTEREST

Artificial Intelligence, Large Language Model, Robotics, Computer Vision, HCI

PUBLICATIONS

Boran Zhao, Haiming Zhai, Haoruo Zhang, Wenzhe Zhao, "LAMMA: A Latency-Aware Design Space Exploration Framework for Multi-CNN on Multi-Core Accelerator", IEEE TCAD (review pending)

RESEARCH EXPERIENCE

University of California, San Diego 06/2023 – 06/2024
Student Researcher, Supervisor: Prof. Sorin Lerner CA, USA

Project: Crammer: Retrieval Augmented Generation for Lecture Transcripts

- Designed and developed an LLM powered RAG system to intelligently query and summarize relevant concepts from large corpus of online lecture transcripts
- Used OpenAI's Whisper to transcribe local video, and Qdrant, a vector database, to embed transcripts for query
- User queries were put into Qdrant to search for relevant transcripts, which were then feed to OpenAI's GPT-3.5-turbo-0125 for information extraction, reorganization and generation

Student Researcher, Supervisor: Prof. Ndapa Nakashole

Reproduction Report: Knowledge-Aware Code Generation with Large Language Models

- Reproduced the published paper KareCoder to evaluate GPT-3.5-turbo-0125's capability of solving competition coding problems
- In KareCoder's framework, the LLM plays two roles: the prompt engineer who generates knowledge-aware prompts upon receiving the problem, the example input/out, and selected portion of built-in knowledge library based on problem categories; and the coder who generates Python code based on the prompt
- Reproduction result showed reasonable performance fluctuation in Pass@k metrics compared to the original

work

Student Researcher, Supervisor: Prof. Manmohan Chandraker

Mini Group Project: Storyboards with Diffusion Models

- Explored the possibility of generating coherent image sequences with diffusion models using prompt engineering and keyframe interpolation
- With a few prompts describing scenes and actions like a script, Stable Diffusion 2 was used to generate high quality key frames. Hard prompt engineering techniques were used to maintain consistent environments and image styles
- Then, a modified OpenAI unCLIP model, Karlo, was used to generate intermediate frames by interpolating keyframes' CLIP embeddings

Tsinghua University 07/2023 – 10/2023

Research Assistant, Supervisor: Prof. Shuguang Li Beijing, China

Project: Aggregation Swarm Robots Inspired by Emergent Properties

- Independently applied swarm robots and reinforcement learning to explore the aggregation process of Dictyostelium discoideum (an amoeboid cellular slime mold)
- Simulated and built swarm robots, and filmed Dicty's aggregation with dark field microscopy
- Analyzed various features that describe the aggregation behavior in both worlds with OpenCV and explored potential similarities and differences between swarm robots and amoebas

Xi'an Jiaotong University 03/2023 - 08/2023

Research Assistant, Supervisor: Prof. Pengju Ren Xi'an, China

Project: LAMMA: A Latency-Aware Design Space Exploration Framework for Multi-CNN on Multi-Core Accelerator

- Proposed a design framework that dynamically allocates computation nodes among multiple CNN inference tasks in run-time
- Constructed innovative methods to support task flow interrupt, which can reallocate occupied computing resources to tasks with higher priority, contributing to the higher probability of meeting real-time deadlines

Boran Zhao, Haiming Zhai, Haoruo Zhang, Wenzhe Zhao, "LAMMA: A Latency-Aware Design Space Exploration Framework for Multi-CNN on Multi-Core Accelerator", IEEE TCAD (review pending)

University of California, San Diego 03/2023 - 06/2023

Team Leader, Supervisor: Prof. Ochoa LA, USA

Project: Deep Online Video Stabilization

- Summarized the latest literature and brainstormed with team to select one related published paper that proposed a deep neural network approach of online video stabilization using Siamese ConvNets
- Reproduced the result and explored the network which uses three types of losses: stability loss (which matches locations of pixels and feature points), shape-preserving loss (avoid distortion of warp grids), and temporal loss (enforce coherency between video frames)

Carnegie Mellon University 06/2021 - 12/2021

Research Assistant, Supervisor: Prof. Min Xu PA, USA

Project: Developing Saliency Detection DNNs for Cyro Tomography

- Developed an unsupervised saliency detection network for cryoET tomographs utilizing modified convolutional U-net, 3DAttention, and other various techniques
- Concentrated on testing different methods of image processing techniques with OpenCV and NumPy, and

prototyping U-net based architectures with PyTorch

Xi'an Jiaotong University 06/2020 - 09/2020

Research Assistant, Supervisor: Prof. Buyue Qian Xi'an, China

Project: Analyzing Multi-modal Electronic Health Records

- Engaged in exploratory research on predicting patients' prognosis based on patients' health record
- Effectively cleaned, filtered, and normalized the patients' medical records, and analyzed the various modalities (vital signs, notes, interventions, and etc.)
- Independently researched Deep Representation Learning approaches and applied creative methods to optimize the results

PROFESSIONAL EXPERIENCE

Inspur Group 06/2019 - 08/2019

Research Assistant Intern, Department of Cloud Computing Beijing, China

- Conducted in-depth research on building a recommendation system for online shopping platforms and compiled related survey reports for weekly meeting with other team members
- Independently modeled and analyzed the systems in the simulation environment with (un)supervised learning, data augmentation, feature extraction, and scalable deployment of neural networks on distributed servers
- Effectively collaborated with other departments to bring out the best of everyone's strengths and conduct accurate information flow synthesis

SKILLS

- **Programming Languages:** Python, MySQL, Java, C, Go
- **Systems & Tools:** Windows, Debian-based Linux, Git, ROS, Docker, Jupyter Notebook, LaTeX
- **Languages:** Chinese (native), English (proficient), Japanese (fluent)