JUSTIN YI

San Francisco, California ⋄ joostinyi00@gmail.com ⋄ 909-342-3421

University of California, Los Angeles 2022 — Computer Science B.S., Mathematics Minor

WORK EXPERIENCE

Software Engineer — **Baseten** — **September 2022 - Present**

- Served as a founding engineer of the **Model Performance** team, implementing various inference optimization strategies like **speculative decoding**, quantization, etc., collaborating with stakeholders from core platform, infrastructure, and GTM teams.
- Designed an **LLM optimization and deployment pipeline** with **TensorRT-LLM** and in-house model weight distribution system for performant inference servers, resulting in **60% greater throughput and 35% cost reduction** for customers serving production traffic.
- Maintained the ML containerization and serving framework **Truss**, adding support for a CLI live reload experience with deployed services and more expressive containerization support.

Data Science Intern — **AI Camp (Edtech Startup)** — **June 2021 - August 2021**

- Project managed student developer NLP projects by defining success criteria and deliverables for machine learning web applications.
- Presented machine learning concepts to hundreds of students nationwide, developing the company's first ML fairness course offering.

Research Assistant — **Pilon Group, UCLA** — **April 2019 - June 2020**

- Trained and evaluated an **attentive generative adversarial network** for image restoration of rain streak distorted images for applications in autonomous driving systems using in house created datasets of 10,000+ samples.
- Performed reverse osmosis of fracking wastewater for CO_2 adsorption for a carbon negative concrete synthesis process.

Research Assistant — **Bhandari Group, Cal Poly Pomona** — **June 2017 - August 2017**

- Studied and implemented methods for autonomous drone navigation in GPS denied environments using OpenCV and Caffe frameworks for Hector SLAM mapping. [poster]

SELECTED PROJECTS

- **On the Complexity and Convergence of Approximate Policy Iteration Schemes**: Literature survey of approximation methods of **Policy Iteration** for **Markov Decision Processes** to with considerations of algorithmic complexity bound analysis, convergence guarantees, and rates of convergence. [poster]
- **Graph Neural Network Projects**: Presented and demonstrated findings of a novel graph convolutional policy network for goal-directed molecular graph generation. Literature survey of GNN applications in the field of programming languages, namely in bug detection, similarity analysis, program synthesis, etc.

LEADERSHIP ACTIVITIES

ACM AI President — **ACM at UCLA** — **November 2019 - June 2022**

- Led 4 committees of 30 members through various workshop, guided project, event, and outreach offerings to the UCLA and surrounding communities.
- Developed and presented multiple 10 week workshops to teach machine learning fundamentals to cohorts undergraduate students – topics included **Neural Networks**, **Deep Learning**, Convolutional NN, Recurrent NN, Fair ML.
- Authored and edited tech policy blogs exploring various relevant socially impactful tech topics: **AI Governance**, Big Tech Regulation, Climate Tech.

Skills: Python, C++, bash, PyTorch, Numpy, Kubernetes, SQL, React, Typescript, LaTex, Ableton