Yixuan Zhou

🏠 1025 Stewart Drive, Sunnyvale, CA, 94085
✉ yixzzhou@gmail.com ☎ (858) 414-3830 💻 Personal Website GitHub LinkedIn

Education

University of California, San Diego (*Summa Cum Laude*) Sept. 2018 - June 2022

- **Double major** in *Computer Science* and *Mathematics – Probability and Statistics*.
- Major GPA in **4.0/4.0**, Overall GPA: **3.98/4.0**.

Experiences

Nuro Inc. *Software Engineer - ML Infra* Mountain View, CA
Python, C++, TensorRT, Keras, Pytorch, Bazel Mar. 2024 - Present

- Maintained and developed new features for a custom ML model compiler and runner, **enhancing inference speed** for large-scale machine learning models deployed on autonomous vehicles with limited computational power.
- Collaborated closely with model engineers to design and implement a staggered execution strategy, resulting in a **30% reduction in amortized latency**.
- Integrated TRT-modelopt to enable **quantization** to optimize models deployed on GPUs specialized in lower precision calculations; Built infra to perform PTQ & QAT, and developed toolings to debug accuracy regression.

Square Inc. *Software Engineer - Payment Onboarding* Seattle, WA (Remote)
Java, Ruby, JavaScript, Ember, Bazel, Terraform Nov. 2023 - Mar. 2024

- Designed and implemented new features in the registration workflow for CA merchants to fulfill KYB regulations.

Stripe Inc. *Software Engineer - Data Privacy Technology Team* Seattle, WA
Java, Python, Go, JavaScript, Dagger, React, Bazel, Trino, Terraform, Proto Aug. 2022 - Nov. 2023

- Designed and implemented an **ML-based workflow** for auto-infer metadata in Stripe data ecosystem. Integrated this with Trino for **attribute-based access control**, enabling data obfuscation during queries.
- **Collaborated with cross-functional teams** to build a web app and CLI tools for dataset annotation, addressing adoption challenges and ensuring the applicability of the annotation framework across diverse datasets.
- Implemented end-user data deletion requests auto-triaging, saving on average **$1,500 per request**.

Award

UCSD Physical Sciences Dean's Undergraduate Award for Excellence [link] Nov. 2021

- 1 of 26 recipients recognized for academic excellence out of 4,000 undergraduates in physical sciences department.

Projects

Quantization of Neural Networks *Undergraduate Researcher* San Diego, CA
Python, Pytorch, Numpy, AWS Sept. 2021 - June 2022

- Designed and implemented the GPFQ post-training quantization algorithm achieving significant size reduction and inference speed acceleration for compressing state-of-the-art neural networks while retaining accuracy.
- **Publication**: Co-authored paper Post-training Quantization for Neural Networks with Provable Guarantees

Project Lim[b]itless *Undergraduate Researcher - UCSD Center of Human Frontiers Lab* Oct. 2020 - June 2022
Python, Pytorch, Numpy

- Employed transfer learning to repurpose DeepLabV3 neural network to segment amputees' limbs from raw images.
- Co-authored TechArxiv Preprint: ImageTransfer Learning with DeepLabv3 to Facilitate Photogrammetric Limb Scanning.

UCSD Math Honor Program *Undergraduate Researcher* Jan. 2021 - June 2022
Python, Polynomial Optimization

- Investigated linear systems' stability (measured by condition number) in certifying polynomial nonnegativity, conducting numerical experiments to test the condition number across different polynomial bases.
- Author of **Honor's Thesis**: Numerics for Different Bases in Certifying Nonnegativity of Polynomials.

Skills

Framework and Technologies: Pytorch, TensorRT, Huggingface, Pandas, Dagger, gRPC, React, Ember, AWS.
Others: English, Mandarin, Surfing, Ultimate Frisbee!