Arsh Zahed

Machine Learning Engineer

Objective: $\max \mathbb{E}\left[\|\text{Experience}\|^2 + \|\text{Knowledge}\|^2\right]$

azahed98.github.io | azahed98 | arsh-zahed | ********** | **********

Experience

TOGETHER AI | AI Engineer
Engineering | July '23 – Current

- Built distributed systems with 6000+ GPUs for inference and training with PyTorch, DeepSpeed, FSDP, Golang, Slurm and Kubernetes.
- Developed API training platform for 30+ LLMs with up to 256B tokens.
- Research in parallel speculative decoding - upto 40% improvement.

TIKTOK | Research Engineer
Speech Audio Music Intelligence | April '22 – April '23

- Developed zero-shot voice design models for age/gender interpolation.
- Trained 4 models with 6 datasets, 2 languages, and 40k speakers.
- Models used by >5 million users in over 160 countries.

NVIDIA | Deep Learning Engineer
AI Applications | July '20 – March '22

- Built Riva model conversion tool to optimize models with Triton, ONNX and TensorRT. Supports 15 pipelines, and accelerates by >12x.
- Built TAO-LM, a language-model tool, used by over 100 customers.

BERKELEY AI RESEARCH | Researcher & Grader
AutoLab | Jan '19 - Jan '20

- Research in Reinforcement, Imitation and Online Learning.
- Reduced failure of safety using uncertainty estimation by 14%.

GOOGLE | Software Engineer Intern
Chrome Media Audio | May '18- Aug '18

- Created TF Estimators experimentation framework to predict the speech coding quality of WaveNet/Lyra while reducing bitrate by 50%.
- Collected 7000 user-rated generations from 3 generative models.

Publications

"On-Policy Imitation Learning from an Improving Supervisor"

- Conference on Robot Learning (CORL), 2019
- Real World Sequential Decision Making Workshop at ICML, 2019.

Projects

ROBUST QA WITH DATA AUGMENTATION AND SSL PyTorch | 2022

- Improved QA F1 and EM by 10% on OOD data via data augmentation and a novel inference-time self-supervised finetuning method.

UNCERTAINTY AWARE PHYSICS ESTIMATION PyTorch | 2021

- Used uncertainty estimation to create an active learning framework for physics estimation. Achieved a >50% decrease in required data.

EXPRESSIVE TTS FROM INFERRED EMBEDDINGS PyTorch | 2020

- Inferred style-embeddings from text to improve generated speech.
- Improved F0 Frame Error by 8% with audible improvement.

METAL - MAML EXPLORATION WITH METRICS TensorFlow | 2019

- Developed Policy Metrics that help guide task-specific exploration.
- Used with imitation learning for 22% reduction in training speed.

Skills

TOPICS & FIELDS
Deep Learning • Generative AI • Speech Processing • Computational Music • Natural Language Processing • Reinforcement Learning

PROGRAMMING
Python • C • C ++ • JavaScript • R • Java • Protobuf • Bash • LaTeX

LIBRARIES & TOOLS
PyTorch • TensorFlow • JAX • Triton • AWS • GCP • Docker • Kubernetes

Education

STANFORD UNIVERSITY
Non-Degree | Sep '21 - Mar '22
Computer Science

UC BERKELEY
B.S. | Aug '16 - May '20
Electrical Engineering & Computer Science

Coursework

COURSERA
Google Cloud Machine Learning Engineer

STANFORD

CS 224n	Natural Language Processing
CS 236	Deep Generative Models

UC BERKELEY

CS 191	Computational Photography
CS 189	Machine Learning
CS 188	Artificial Intelligence
CS 170	Algorithms
CS 162	Operating Systems
CS 161	Computer Security
EE 225b	Digital Image Processing
EE 127	Convex Optimization
EE 126	Probability Theory
EE 123	Discrete Signal Processing
EE 120	Signals & Systems
Math 141	Differential Topology
Math 110	Linear Algebra
Math 104	Real Analysis
Music 108	Music Cognition
Stat 154	Stochastic Processes
Stat 153	Time Series Analysis