Billy Zhong

https://billyzh.ong

(925) 785-4285
billyczhong@gmail.com
https://github.com/BillyZhong

Work

Meta (Modern Recommendation Systems) — *Software Engineer* 2022-Present

Contributed to higher GPU efficiency and data throughput within multiple recommendation models via optimizations at various levels, saving GPU capacity and cost

Enabled several serving-time optimizations (hardware scheduling, graph compilation, quantization lowering) for the T2I recommendation model, producing an end-to-end speedup of 30% during inference

Horizontally fused several dense modules within the VDDv4 recommendation model at the PyTorch level to produce an end-to-end speedup of 4% during training and 7% during inference

Leveraged OpenAI's Triton language to write kernel code to vertically fuse the RMS Norm operation, resulting in an 80% speedup and a 50% reduced GPU peak memory usage at the operator level

Twitter — *Software Engineering Intern* 2021

Designed and deployed retry pipeline and dead letter queue system using Java and Kafka for ad impression events that failed initial processing

Recovered 1% of ad revenue resulting from failed events and analyzed repeatedly failing events to identify potential system improvements

DeepMap — *Computer Vision Intern* 2019

Researched and developed statistical methods to automate road lane line feature labeling within satellite imagery

Implemented these methods into tools to assist with manual feature annotation of high-definition maps

Achieved an accuracy of 90%, decreasing manual annotation time and error

Projects

Self-Balancing Text String Trees 2021–2022

Worked under Professor Michael Fisher for Yale senior research project to augment self-balancing binary search trees with linked-list string representations

Designed novel data structure that allows modifying strings and tracking characters across modifications with logarithmic algorithm runtime

Implemented new data structure and algorithms in C++ library

Chinese Study Tool 2019

Developed computer vision application using template matching to recognize Chinese characters in digital textbook pages to accelerate study while taking Chinese classes at Yale

Created automated annotation tool to display pronunciations and translations of characters

Dynosaur 2016-2018

Initiated machine learning project for Biology class to utilize neuroevolution to create an autonomous bot to play the Google Dinosaur Runner Game

Designed an asynchronous, parallelized neuroevolution algorithm that could produce a bot capable of scoring 500 within an hour

Created interactive web dashboard to visualize model performance and bot progress

Education

Yale University — *Computer Science (B.S.) and Mathematics (B.A.)* 2018–2022

CPSC 223 — Data Structures
CPSC 366 — Intensive Algorithms
CPSC 460 — Automata Theory
CPSC 465 — Theory of Distributed Systems
CPSC 468 — Computational Complexity
PHIL 267 — Mathematical Logic
PHIL 427 — Computability and Logic
PHIL 439 — Modal Logic
CPSC 323 — Systems Programming
CPSC 413 — Computer System Security
CPSC 447 — Quantum Computing
CPSC 452 — Deep Learning
CPSC 467 — Cryptography
CPSC 470 — Artificial Intelligence
CPSC 475 — Computer Vision
CPSC 476 — Advanced Computer Vision
MATH 230 — Vector Calculus and Linear Algebra
MATH 244 — Discrete Mathematics
MATH 270 — Set Theory
MATH 305 — Real Analysis
MATH 310 — Complex Analysis
MATH 350 — Abstract Algebra
MATH 354 — Number Theory
ECON 351 — Mathematical Game Theory

Skills

Python, PyTorch, Triton (Language), C/C++, Java, Javascript, OpenCV, Kafka, SQL, Git, Bash, LaTeX