Pawan Jayakumar

github Website email

Education

University of California San Diego Sept 2024 - Present
Master of Science in Computer Science *GPA: 4.0/4.0*

University of Virginia Aug 2020 - May 2024
Bachelor of Science in Computer Science *GPA: 3.83/4.0*

Thomas Jefferson High school for Science and Technology Aug 2016 - May 2020

Coursework

Software Engineering, Data Structures and Algorithm Design, Operating Systems, Machine Learning, Parallel Processing, Hardware Accelerators, Robotics, Probability theory, Linear Algebra

Experience

Pytorch | *Open Source Software Engineer* May 2024 - Sept 2024

- Actively engaged in the development of TorchAO, a library for performing architecture optimization for AI model inference and training by opening issues, performing code reviews, and updating documentation
- Created a new data type for low-bit quantization using tensor sub-classing and bit-packing to reduce memory cost of network weights by 2-4x
- Implemented Activation-aware Weight Quantization (AWQ) which is used by over 3400 models on Huggingface

Capital One | *Software Engineering Intern* June 2023 - Aug 2023

- Built and deployed a full-stack cloud application using React, GraphQL, and AWS Dynamo DB, which is used by over 15,000 monthly associates
- Optimized local development build times by decoupling our service, saving 100+ hours of development time

Capital One | *Software Engineering Intern* Jun 2022 - Aug 2022

- Designed and engineered a full-stack cloud application to track and display changes in vulnerability reports to Capital One associates using Angular, and a variety of AWS database management services
- Negotiated with the product team, presented design choices that would improve customer experience, performed code reviews, and proactively asked for feedback

University of Virginia | *Teaching Assistant* Aug 2022 - Dec 2022

- Led 100+ students in laboratory sessions and office hours by conducting code reviews and peer mentoring

Research projects

LLM Reasoning Research Jan 2024 - Present

- Fine-tuned modern LLM's to generate sentence level embeddings from chain of thought reasoning data
- Currently pre-training various auto-regressive and diffusion models to perform next sentence generation

Temporal Downsampling for Byte-Transformers Sep 2024 - Dec 2024

- Improved the accuracy of BERT-style byte level transformer by 30% on speech transcript classification benchmark using sequence dimension down sampling with convolutions
- Outperformed subword-tokenizer methods when text contained misspelled words (improved robustness)

Policy Evaluation Benchmark Feb 2024 - August 2024

- Constructed a testing harness for policy evaluation algorithms such as ROS and BPS
- Parallelized model training and inference on compute clusters using Slurm and Weights and Biases

Slider Mar 2022 - Mar 2023

- Co-developed and published an award winning puzzle game called Slider which has over 10,000 unique players