Pawan Jayakumar

github Website ✉ pawan.jayakumar@gmail.com

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, Databases, Distributed Systems, Hardware Accelerators, Robotics, Probability Theory, Linear Algebra

Experience

Pytorch | *Open Source Software Engineer* May 2024 - Sept 2024
- Engaged in the development of TorchAO, an architecture optimization library for AI model inference, by opening issues, performing code reviews, and updating documentation
- Implemented Activation-aware Weight Quantization (AWQ) which is used by thousands of models on Huggingface

Capital One | *Software Engineering Intern* Summer 2023 + 2024
- Built and deployed a scalable full-stack cloud application using React, GraphQL, and AWS Dynamo DB
- Optimized local development build times by decoupling our service, saving 100+ hours of development time
- 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

Other Projects

LLM Security Jan 2025 - April 2025
- Uncovered a vulnerability in OpenAI's deep research tool which allowed for the discovery of exposed API keys
- Applied GCG attacks onto DeepSeek distilled reasoning models showing that test time inference doesn't inherently improve adversarial defenses
- Reproduced emergent misalignment on Gemini-Flash-1.5 which showcased harmful behavior 2.5% of the time when using prompt templates

Mix Lab | *Researcher* Jan 2025 - Present
- Fine-tuned language models to create auto encoders for sentence level embeddings
- Currently speeding up video diffusion models through one step generation distillation

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)

Slider Mar 2022 - Mar 2023
- Co-developed and published an award winning puzzle game called Slider which has over 10,000 unique players

Skills

Languages: Python, C/C++, CUDA, Triton, Bash, SQL, C#, JavaScript, HTML, CSS
Tools: Github, Docker, AWS, JIRA, Weights and Biases, Llama.cpp
Frameworks: PyTorch, MPI, NCCL, React, Angular, Rest, GraphQL, Tailwind