Rayan Singh

https://rayan.love

Email : rayanpurisingh@gmail.com
Mobile : +1-(650)-960-5056

Education

University of Illinois Urbana-Champaign Urbana, IL
Bachelor of Science, Computer Engineering, Mathematics Minor *Aug. 2021 – May. 2025*

- Relevant Coursework: Introduction to Algorithms and Models of Computation, Computer Systems Engineering, Data Structures, Linear Algebra with Computational Applications, Abstract Linear Algebra, Multivariate Calculus, Computer Systems and Programming, Artificial Intelligence, Programming Languages and Compilers

Experience

AMD San Jose, CA
AI/ML Software Engineer Intern *May. 2024 - Present*

- Deployed and validated state-of-the-art Vision Transformer model on AMD's flagship AI Engine NPU Architecture
- Reformulated CUDA-dependent fusion transformers as ONNX models for deployment to AI Engine NPU
- Refactored convolution layers in large multi-modal fusion transformers for cache-efficient data movement patterns

AMD San Jose, CA
AI/ML Intern *May. 2023 - Aug. 2023*

- Modified the Microsoft ONNX Runtime framework to integrate a novel 24-bit ”rotating point” quantization format
- Devised algorithms for the conversion and quantization of new quantization scheme in C and Intel Intrinsics
- Demonstrated the effectiveness of the new quantization scheme through validation of 96% maintained inferential accuracy, hypothesized 5x speedup on Resnet50, MobilenetV2, and Microsoft Win24 models
- Conducted research on quantization calibration techniques and their impact on maintaining inferential accuracy
- Built testing library for Computer Vision quantization and accuracy using Microsoft Olive and ONNX Runtime

Webscale Networks Boulder, CO
Software Intern/Researcher *May. 2022 - Aug. 2022*

- Devised heuristic that leverages statistical models and Neural Network predictions for accurate anomaly detection
- Developed a specialized tool utilizing GoLang and Apache ECharts to efficiently identify and label discords within real-time data streams from online stores

DesCyPhy Lab Los Angeles, CA
Research Assistant, University of Southern California *June. 2019 - Aug. 2019*

- Conducted research of emerging technologies aimed at countering the impact of SAT attacks, attacks in which a SAT solver is exploited to sniff encrypted circuit keys

Projects

ML-Autonomous Driving Project on Ultra96 Board: Object-detection demonstration of a ZZSoC board to detect objects such as animals and people
Operating System: Built Linux-inspired operating system for Intel x86 ISA using C and x86 assembly. Includes support for program execution, memory paging, scheduling, and the EXT2 file system. Includes custom user program for audio recording and playback, as well as MIDI file support

Technical Skills

Languages: C, C++, Python, Java, x86 Assembly, SystemVerilog, LaTeX, OCaml
Frameworks: ONNX, ONNXRuntime, PyTorch, TorchVision, HuggingFace Transformers, Microsoft Olive
Developer Tools: Git, Docker, Conda, Mamba, Visual Studio Code

Awards and Other Activities

1st Place, Operating System Competition: Designed and implemented x86 operating system in eight weeks and won best operating system kernel out of 80 teams. Hosted by the UIUC ECE department.
2nd Place, FPGA Design Showcase: Implemented 8-track audio sequencer software in C with 6 drum tracks and 2 square wave synthesizers. Designed MicroBlaze SoC and custom AXI audio engine IP ”Rayan's Sound Machine” in SystemVerilog. Worked solo and came 2nd place out of 140 teams of two. Hosted by the UIUC ECE Department.
Chess: Mentored by Indian Grandmaster; 1st place in a tournament with 1300 participants of all skill levels