Avi Singhal

avisinghal6@gmail.com|9452491966 |linkedin.com/in/avi-singhal99|Fremont, California|github.com/avisinghal6 |avisinghal.com

EDUCATION

Rice University, Houston Aug 2022 - Dec 2023
Master of Computer Science, GPA- 4.0/4.0 Texas, USA
Coursework: Probabilistic Algorithms and Data Structures, Deep Learning for Vision and Language, Machine Learning, Machine Learning with Graphs, Design and Analysis of Algorithms, Software Engineering Methodology, Computer Architecture

Delhi Technological University Aug 2017 - Jun 2021
Bachelor of Technology- Electronics and Communication, GPA- 8.75/10 New Delhi,
India Coursework: Microprocessors and Interfacing, Embedded Systems, Pattern Recognition, Web Development

SKILLS

Programming Languages: Python, C++, Java, SQL, JavaScript, Matlab, Spark, CUDA
Technologies and Frameworks: PyTorch, TensorFlow, Hugging Face, AWS, GCP, ONNX, Docker, Linux, Langchain

WORK EXPERIENCE AND INTERNSHIPS

TetraMem Inc.: Machine Learning Model Development Engineer Feb 2024 - Present
Skills: Python, C++, PyTorch, TensorFlow, Hugging Face, ONNX, AWS, Docker, Linux, Git California, USA

- Leading research & development of efficient edge integer(UINT8) inference framework of LLMs & CNN models with **<1%** quantization loss. Detected & solved critical bugs in the quantization framework leading to accuracy increase from **0.5% to 90%** for vision transformers, obtained correlation greater than **0.99** for all layers and models in the model zoo. Paving the path for 4-bit quantization of ViTs and LLMs.
- Prototyping & fine tuning vision transformers in distributed settings in AWS, compressing state space-based MAMBA models. Developing architectures and search spaces for audio applications, object, and face detection.
- Developing, training models and demos for face and gaze tracking for AR/VR applications.
- Developing search algorithms to facilitate switching between float/integer and mixed precision execution based on layer outputs & performance and increase flexibility.
- Developing and benchmarking tiny stories LLMs, identifying high performing architectures and schemes for execution on custom AI accelerators (in memory compute).

TetraMem Inc.: Software/ML Intern May 2023 - Nov 2023
Skills: Python, C++, PyTorch, TensorFlow, Hugging Face, ONNX, AWS, Docker, Linux, Git California, USA

- Achieved high accuracy **(~85%)**, low quantization (**uint8**) loss **(~1%)**, low latency **(<5K** MAC operations), compressed model size**(<350KB**) by building neural architecture search and post training quantization framework for edge devices.
- Improved accuracy by **~3%** to reach **~88%** accuracy for CIFAR10 on resource constrained devices using **joint optimization** of NAS & Hyperparameter optimization (HPO) inspired by CVPR 23's MA2ML with reinforcement learning.
- Introduced support for **10+** intricate ONNX operators and simulation of noise to ML compiler to enhance model inference on AI accelerators and **improve** accuracy by **at least ~5%.**

Texas Instruments (TI): Test Engineer Jan 2021- Jul 2022
Skills: Python, C++, git Bangalore, India

- Reviewed large C++ code base, crafted scalable and efficient test program for production release. Resolved **50+** bugs, incorporated **20+** features to **enhance debugging** in the verification tool crafted at TI, **recognized** as top contributor.
- Constructed a parasitic extraction tool in python which **reduced** test hardware redesign **time**, **cost** by **30%**.

RELEVANT PROJECTS

Adaptive Learning with Dynamic Batch Creation Using Near-Neighbors Aug 2022 - May 2023

- Created an adaptive batch creation algorithm to create new batches using near neighbors of samples with highest gradients from previous batch like the paper from ICLR 2016. Attained **~5%** faster convergence compared to random batches.
- Conducted extensive experiments with several thresholds & variations, monitored experiments using weights and biases.

Mechanistic Interpretability of Transformers Aug 2022 - Dec 2022

- Trained a decoder only transformer with only attention layers from scratch inspired by ChatGPT. Analyzed attention scores via heatmaps, revealing insights including copying mechanism, skip gram behavior in ~**30%** attention heads.
- Performed extensive debugging and parameter tuning to achieve successful training of the transformer model.