Avi Singhal

as278@rice.edu |+19452491966 |linkedin.com/in/avi-singhal99|Houston, Texas, USA |github.com/avisinghal6 |avisinghal.com

EDUCATION

Rice University, Houston, Texas, U.S.A Aug 2022 - Dec 2023
Master of Computer Science (MCS), GPA- 4.0/4.0
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

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

SKILLS

Programming Languages: C++, Java, Python, JavaScript, HTML, CSS, Matlab, SQL.
Technologies: Applied Machine Learning, Deep Learning, CNN, RNN, LSTM, Cyber Security, NLP, Computer Vision, OOPs, Generative AI, Reinforcement Learning (RL), Transformers, GNNs, MERN stack.
Frameworks: Express.js, MongoDB, React.js, Node.js, Firebase, Redis, PyTorch, TensorFlow, ONNX, AWS, Git, OpenCV

WORK EXPERIENCE AND INTERNSHIPS

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

- Achieved high accuracy **(~85%)**, low quantization (**uint8**) loss **(~0.5%)**, low latency **(<5K** MAC operations), small model size**(<350KB**) by building neural architecture search (NAS) and post training quantization framework for **resource constrained** devices for computer vision.
- Improved accuracy by **~3%** to reach **~88%** accuracy for CIFAR10 on edge devices using **joint optimization** of NAS and Hyperparameter optimization (HPO) inspired by CVPR 23's MA2ML using RL.
- Model development, research & implementation spanning computer vision, AI ISP, and audio applications for edge devices.
- Introduced support for **5+** intricate ONNX operators with unit tests and simulation of noise to ML compiler to enhance model inference on AI accelerator and **improve** accuracy by **at least ~3%.**
- Building framework to support quantization and inference of transformer-based models for edge devices starting by researching and prototyping quantized Efficient Former, paving the path for LLM support on TetraMem AI accelerator.

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

- Reviewed large C++ code base, crafting scalable and efficient test program for production release. Resolved **50+** bugs, incorporated **20+** features for **enhanced debugging** to the verification tool crafted at TI, resulting in **recognition** as best user.
- Constructed a parasitic extraction tool with user interface using python scripts. The tool helped **reduce** test hardware redesign **time**, **cost** by **30%** and better correlation of simulation output with tester results. Work published at TI conference.

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.

Graph Based Recommender System Jan 2023 - May 2023

- Performed comparative analysis based on scalability, precision, latency on YouTube dataset using: Pixie Random Walk (PRW), Random walk based embeddings and link prediction without GNN with the latter outperforming all methods.

Mechanistic Interpretability of Transformers Aug 2022 - Dec 2022

- Built and trained a decoder only transformer inspired by ChatGPT, from scratch with only attention layers. Analyzed attention scores via heatmaps, revealing insights including copying mechanism, skip gram behavior in ~**30%** attention heads.

Recipe Generation from Videos Jan 2023 - May 2023

- Engineered a recipe generation pipeline leveraging pretrained models. Executed parallel dense video captioning for event proposals, reduced redundancy with cosine similarity and elevated image analysis of frames using YOLO, harnessed BLIP for image captioning, Resnet for frame-to-feature conversion. Employed ChatGPT for concise caption summarization.
- Evaluated performance with BERTScore, obtaining commendable results despite absence of model training with latency ~**5s**.

Video Conferencing Application Jul 2023 - Aug 2023

- Designed a video conferencing application using React.js for front-end and Express.js for backend using Socket.io, WebRTC. Incorporated functionality for audio muting, disabling video, screen sharing, and **5+** unique features being added.

Social Media Website Jan 2022 - Jul 2022

- Developed social media app using MERN stack, HTML, CSS. Implemented authentication using multiple strategies from passport.js. Incorporated creation, deletion of posts, comments, likes, friend requests and personal real-time chat rooms.