Avi Singhal

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

EDUCATION

Rice University, Houston, Texas, U.S.A 08/2022-12/2023
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 08/2017-06/2021
B.Tech- Electronics and Communication , GPA- 8.75/10
Coursework: Computer Architecture, Microprocessors and Interfacing, Embedded Systems, Pattern Recognition, Web Development

SKILLS

Programming/Scripting Languages: C++, Java, Python, JavaScript, HTML, CSS, MATLAB, SQL.
Technologies: Applied Machine Learning, Deep Learning, CNN, RNN, LSTMS, Cyber Security, NLP, Computer Vision, OOPs, Generative AI, Reinforcement Learning (RL), Transformers, GNNs, MERN stack.
Software Frameworks and Tools: Express.js, MongoDb, React.js, Node.js, Firebase, Redis, Pytorch, Tensorflow, ONNX, AWS.
Additional Skills: Linear Algebra, Calculus, Creativity, Collaborative mindset, Agile, Growth-minded

WORK EXPERIENCE AND INTERNSHIPS

TetraMem Inc.: Software/ML Intern 05/2023-12/2023
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 and implementation spanning computer vision, AI ISP and audio applications for edge devices.
- Added 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 with quantized Efficient Former, paving the path for LLM support on Tetramem AI accelerator.

Texas Instruments (TI): Test Engineer 07/2021-07/2022
Skills: Python, C++, git Bangalore, India

- Reviewed large C++ code base, designed scalable and efficient test program for production release. Resolved bugs, incorporated **20+** features for **enhanced debugging** to verification tool developed at TI, resulting in **recognition** as best user.
- Developed 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 08/2022-05/2023

- Created new batches using near neighbors of samples with highest gradients similar to the paper from ICLR 2016. Achieved faster train/val convergence compared to random batches for convex cases. Verified approach on datasets from Liblinear.

Graph Based Recommender System 01/2023-05/2023

- 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 08/2022-12/2022

- Built and trained a decoder only transformer inspired by Chatgpt, from scratch with only attention layers. Plotted attention scores of keys and queries using heatmaps, observed some copying mechanism, skip gram behavior in a few attention heads.

Recipe Generation from Videos 01/2023-5/2023

- Built a recipe generation pipeline leveraging pretrained models. Implemented parallel dense video captioning for event proposals, reduced redundancy with cosine similarity. Enhanced the frames using YOLO, harnessed BLIP for image captioning, Resnet for frame-to-feature conversion. Employed BARD,Chatgpt for concise caption summarization.
- Evaluated performance using BERTScore, achieved commendable results despite absence of model training.

Video Conferencing Application 07/2023-08/2023

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

Social Media Website 01/2022-07/2022

- Developed social media platform using MERN stack and 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.

OPEN SOURCE CONTRIBUTIONS: Huggingface/accelerate(commits), /transformers(commits), Microsoft NNI(commits)