Sai Vivek Peddi

Objective: Deep Learning Systems

530-220-7864 San Jose, California 95112 @ svpeddi@ucdavis.edu saivivek-peddi Saivivek-Peddi

EXPERIENCE

JP Morgan Chase — Palo Alto, CA
Senior Software Engineer — Jan '23 - Ongoing
- Developing a High-Performance Transaction Processing System Tailored for Corporate and Investment Banking.
- Ensuring System Capability to Handle **4,000 TPS** with a Maximum Delay of **200ms** Per Transaction
- Designing and Implementing Balance Loaders for Transaction Processing Systems using **Spring**, **Redis**, and **Kubernetes**.

KFintech — Hyderabad, India
Solutions Architect — Feb '19 - July '21
- Architected **Digix**, Data visualization tool for **18** AMCs with over **$137B** of AUM and **100M** investor accounts. ***(Architecture)***
- Developed **InPro**, an AML tool for screening investors against **40M** world check records to detect fraud. ***(Architecture)***
- Led a team of **5** to write large scale data transformation Jobs in **Spark**. Reduced query times by **40** fold with new pipelines.
- Implemented a **YOLOv3** model for cropping One Time Mandates (OTMs). Achieved **98% accuracy** for OTM detection.

IIIT Hyderabad — Hyderabad, India
Research Assistant, ML & Speech Processing Lab — Aug '18 - Jan '19
- Developed an **Intrusion Detection System (IDS)** to tag traffic with malicious intent from darknets like **Tor** & **I2P**.
- Experimented with **Bayes Net, C4.5, & RF** for flow level classification with the packet capture (**PCAP**) data from Tranalyzer.

EDUCATION

University of California, Davis — Davis, CA
Masters degree in Computer Science (GPA: 4/4) — Sep '21 - Dec '22
- **Courses:** Adv Deep Learning, ML & Node Discovery, Modern Parallel Architecture, Distributed Database Systems, Quantum Simulations, Adv Visualization, Computer Architecture, Computer Networks, Operating Systems.
- **TA:** ML - *Fall 2021*, *Spring 2022* **,** Programming Languages - *Winter 2022* **,** Design & Analysis of Algorithms - *Summer 2022*

Birla Institute of Technology and Science (BITS), Pilani — Hyderabad, India
Bachelor of Engineering (Hons.) in Electronics & Communication Engineering (GPA: 7.89/10) — Aug '15 - Dec '18

SKILLS

- **Programming:** Python, Java, C, C++, JS, Cuda, SQL. • **ML:** PyTorch, CuDNN, NCCL, ROCm, Spark, Hadoop, Hive, Kafka.
- **Other Skills:** AWS(EMR, Athena, RDS, S3, SageMaker), Terraform, pandas, Mongo, HDFS, Spring Batch, git, docker, K8s.

CERTIFICATIONS

- **AWS Certified Solutions Architect – Associate** : Credential

PROJECT HIGHLIGHTS

LLM Hardware Acceleration Suite — Aug'23 - Ongoing
- Leading a team of **15 CS grad** students in developing hardware accelerators for **LLMs**, in partnership with **AMD** researchers.
- Conducting research on **Heterogeneous Training and Inference**, focusing on cross-vendor GPU compatibility in **ML tasks**.
- Researching **DPU-based Smart NICs** for LLM task offloading, aimed at enhancing network and processor efficiency.

Multi-head attention with Sparse GPU kernels — Oct '22 - Dec'22
- Researched parallelism and sparsity in the **multi-head attention** module inside each encoder of a **Vision Transformer.**
- Developed the CUDA kernels for Self-attention using **cuSparse SPMM** and **cuBLAS SPMM** and experimented on **1660 Ti**.
- Able to achieve a **100x** speedup with **85%** sparsity and **97.5%** compute capacity of the GPU making it arithmetic bound.

Leveraging network delay variability to improve QoE of Latency Critical (LC) services — Jul '22 - Sep '22
- Developing **Kubernetes** modules to schedule and serve requests to attain **End-to-end Service Level Objectives** on cloud.
- Built the **Control Plane** for scaling using **telemetry** info from **Prometheus** and used Grafana to visualize the metrics.
- Able to guarantee a QoE target for more than **75%** utilization using EDF (**15%** higher than FCFS) for every pod in the cluster.

Application of Transformers in Audio Classification (Paper) (Code) — Mar '22 - Jun '22
- Built **self-supervised** models for Audio classification using transformers to tackle **inductive bias** from the existing CNNs.
- Experimented Vanilla **ViT**, **MAE**, **DeIT** and **Swin Transformers** on **Mel-Spectograms** extracted from audio clips.
- Achieved **94.27%** accuracy with **DeIT** outperforming the SOTA **D-CNN(85.14%)** on UrbanSound8k dataset.

Privacy First Query Optimized Horizontal Partitioning using Machine Learning (Code) — Jan '22 - Mar '22
- Developed an ML model using **KModes** to build partitions automatically based on past run queries.
- Achieved up to **3x** speedup using **vertical partitioning** and up to **9x** speedup using **horizontal partitioning**.

Volumetric Isosurface Rendering with Deep Learning-based Super-Resolution (Video) — Sep '21 - Dec '21
- Developed an advanced visualization system to interact with **3D Volumetric Data** sets like CT Scans & Super Novas.
- Reduced the rendering time of a CT Head, with Ambient Occlusion, from **4.2s** to **0.071s** using the super-res network.