Sandhya Rani Narravula

☎ +1-934-246-8787 ✉ narravulsandhyarani@gmail.com Linkedin Github

EDUCATION

Stony Brook University **Stony Brook, NY**
Master of Science in Computer Science *Aug 2024 - Dec 2025*

RMK Engineering College **Tamil Nadu, India**
Bachelor of Engineering in Computer Science and Engineering, CGPA: 9.21/10 *July 2018 - May 2022*

Skills Summary

Languages: Python, C, C++, C#, Java, JavaScript, HTML, CSS, SQL, Bash
Frameworks: PyTorch, TensorFlow, Keras, JAX, ONNX, Hugging Face Transformers, CI/CD
Tools: Git, Docker, EC2, AIMET, MLflow, Weights & Biases, TensorRT, Open3D, Apache Spark, Ray, Dask
Technologies: Deep Learning, Large Language Models (LLMs), Reinforcement Learning (RL), AWS, Kubernetes, Multi-Agent RL (MARL), Model Compression, Transfer Learning, Computer Vision, NLP, Kubernetes, MLOps, Edge AI, Federated Learning, OpenCV, Lambda, NumPy, Pandas, Matplotlib, Azure
Other Skills: Quantization , Pruning, AutoML, Graph Neural Networks (GNNs), Transformer-Based AI, Agile, Retrieval-Augmented Generation, Model Deployment, Performance Benchmarking, containerization

Work Experience

Stony Brook University **Stony Brook, NY**
Research Assistant – AI/ML, Prof. Erez Zadok & Prof. Arie Kaufman *Jan 2025 - Present*

- Conducting research on **AI-driven workload scheduling** for large-scale **LLM training**, in collaboration with **IBM Research**.
- Developing **Reinforcement Learning (RL)**-based **task scheduling algorithms** using **Deep Q-Learning (DQN)** and **Proximal Policy Optimization (PPO)**, optimizing GPU cluster utilization for **40% faster fine-tuning of Large Language Models (LLMs)**.
- Engineering a **Transformer-based resource allocation framework**, enabling **adaptive scaling** and reducing **compute overhead by 25%** in distributed AI training.
- Designing a **self-optimizing ML workload orchestrator** that dynamically schedules AI tasks using **Graph Neural Networks (GNNs)** and **Attention Mechanisms** to **predict optimal task execution sequences**.
- Leveraging **LLM-driven optimization techniques** for cloud-based AI workloads, ensuring efficient scheduling of **multi-modal ML pipelines** across heterogeneous computing environments.

MultiCoreWare Inc. **Chennai, India**
Software Engineer - Machine Learning *Dec 2021 - July 2024*

- Progressed from Project Intern (Dec 2021 - June 2022) to Software Engineer (July 2022 - July 2024).
- Led quantization and optimization of **GKT, BEVDet, BEVFusion, Cavaface, Rankpose, and Img2Pose** models for autonomous vehicles, reducing latency by up to **10x**, boosting inference speed by **40%**, and increasing edge FPS by over **10x**, enabling real-time safety systems, facial recognition, and pose estimation.
- Redesigned deep learning model architectures for **2D and 3D object detection models** (including **BTCDet, Point Pillars, YOLO, and RetinaNet**), reducing parameter count by **30%** while preserving **98%** of original accuracy, improving efficiency and facilitating deployment on resource-constrained edge devices.
- Leveraged **AIMET (AI Model Efficiency Toolkit)** for advanced model quantization, implementing techniques such as **cross-layer equalization, AdaRound, and mixed precision** to convert models from **FP32 to INT8 or FP16**, significantly reducing model size and inference time.
- Demonstrated expertise in **ONNX (Open Neural Network Exchange)**, utilizing it for model interoperability and optimization across different deep learning frameworks, enhancing deployment flexibility and performance.
- Earned the **Customer Delight Award** for optimizing the **GKT model**, achieving a **15% improvement** in overall system performance and a **15x latency reduction** on the **Qualcomm Snapdragon 8cx Gen 3** by implementing advanced quantization and optimization techniques.

Projects

Multi-Agent Reinforcement Learning for Autonomous Vehicles

- Developing **multi-agent reinforcement learning (MARL)** models for **Autonomous Vehicle (AV) platooning and traffic coordination**, reducing **collision risk by 45%** and **improving traffic flow by 30%**.
- Integrating **Graph Neural Networks (GNNs)** and **Transformer-based perception models** to enable **real-time AV communication and cooperative driving strategies**.
- Optimizing **autonomous vehicle navigation** using **deep imitation learning** and **model-based reinforcement learning** for **low-latency decision-making in dynamic environments**.

Alzheimer Detection Using SMLT

- Engineered a **machine learning model** for early Alzheimer's detection, attaining **94.35% accuracy** and potentially curtailing diagnosis time by **40%** compared to traditional methods.
- Devised **data preprocessing techniques** that bolstered model robustness, enabling it to process **30% more diverse patient data** without compromising accuracy.

University Chatbot Using AI

- Designed and deployed an **AI-powered chatbot** that reduced student inquiry response time by **80%**, handling over **1000 queries per day** with a **95% satisfaction rate**.
- Implemented **personalized learning path recommendations**, increasing student engagement by **35%** and improving average test scores by **15%**.