Sai Surya Vamsi Krishna Kocherla

Minneapolis, Unites States | koche156@umn.edu | LinkedIn | GitHub

Education

University of Minnesota - Twin Cities **Minneapolis, MN**
Masters of Science | Robotics Aug 2024 – Aug 2026 (Expected)
Relevant Coursework: *Machine Learning Fundamentals, Robot Vision, Intelligent Robotics, Image Processing & Application*

Skills & Interests

Hardware Skills: Jetson Orin, Jetson AGX, Raspberry Pi, Arduino, Embedded Systems, Drones.

Technical Skills: **PyTorch**, **TensorFlow**, Scikit-Learn, Pandas, **Computer-Vision**, **Machine-Learning**, **Deep-Learning**, Neural-Networks, NLP, **OpenCV**, **CNNs**, **Transformers**, **Visual-Transformers**, RNNs, Stable-Diffusion, **Docker**, **Kubernetes**, **Kubernetes-Operator**, **KNative**, **KServe**, Kubeflow, Airflow, Cert-Manager, **Flask**, GitLab CI/CD, **Linux**, Arch Linux, **Linux File System**, MongoDB, **MySQL**, **PostgreSQL**, ClickHouse, **Redis**, RabbitMQ, **Kafka**, KEDA, **Prometheus**, **Grafana**, **OpenTelemetry**, **EKS**, **AKS**, **GKE**, **Istio**, **Nvidia Tech Stack**, **Triton Inference Server**, GPU Operator, **K3S**, **K3d**, SendGrid, **Git**, **DVC** (Data Version Control), ArgoCD, KeyCloak

Cloud Platforms: GCP, **AWS**, **Azure Cloud**

Programming Languages: Python, **Golang**, **C**, **C++,** Rust, CUDA, JAVA, HTML, NodeJS, Shell Scripting, Helm Charts

Other Skills: Designing Applications Based on Microservice Architecture, Deploying GPU-based high-performance Workloads into Kubernetes clusters, Implementing DL algorithms based on Research Papers, Configuring HPC Systems, Project Management, Process Planning, Leading project teams towards execution, Photography, Adobe Lightroom, Adobe Photoshop, Data Modeling, Agile (Scrum/Kanban), SOLID, Debugging, Root Cause Analysis, Data Modeling,

Work Experience

Upjao Agrotech **Ahmedabad, India**
Senior Computer Vision Scientist **(Jan 2019 – Jul 2024)**

- Applied knowledge of neural network architectures, loss functions, activation functions, and training techniques to **implement research papers**.
- Efficiently **implemented, deployed, and optimized deep learning algorithms** on resource-constrained **edge (Raspberry Pi, Jetson Boards, etc.) and mobile devices**, through model analysis, quantization, and redundancy reduction techniques. **Leveraged onboard GPUs on Android platforms** to achieve **lower inference times** through **Open Source Vulkan APIs, CUDA APIs,** and **PyTorch Mobile**.
- Leveraged cutting-edge **deep learning models and classical computer vision algorithms** to develop robust and accurate computer vision solutions, including object detection, segmentation, and classification, while optimizing model efficiency and training time.
- **Architected, engineered, and deployed** a highly scalable and production-ready Serverless inference platform on **Kubernetes** built using **NVIDIA's Triton inference server**, **KServe, KNative,** and **gRPC protocol,** leveraging a microservice architecture to host and infer multiple deep learning models with low latency of below 500ms.
- Employed **multithreaded programming and structural pipelining** to **enhance inference times by 50%** also **reducing the infrastructure costs by 60%** and leveraged half-precision (**FP16**) and integer precision (**INT8**) models, optimized for FP16-capable GPUs like NVIDIA A100 and Tesla T4, to further accelerate inference speed in production environments.
- **Established automated testing and deployment** procedures for machine learning and deep learning models through the implementation of **CI/CD pipelines**, **incorporating multi-stage pipelines** to guarantee thorough validation of the developed models.
- Developed an in-house **Model Caching algorithm**, which reduced the **scale-up time from 15-20 minutes to just 3-4 seconds.**

Tata Consultancy Services **Hyderabad, India**
AI / ML Developer **(Jun 2021 – Apr 2023)**

- **Selected into TCS Rapid Labs Cohort**, **an elite R&D team** focused on cutting-edge technologies including **AI and Robotics.**
- **Designed and developed NLP-based, complex proofs-of-concept (POCs)** within three months, encompassing **core AI algorithms, backend, and frontend components.**
- Developed an algorithm to **remove and identify complex repeated data** as part of the **data-cleaning process,** improving efficiency by 20%
- Developed **data filtering algorithms** to verify and validate images and logos, extract text from images, remove and identify complex repeated data, and localize and extract table data embedded in textual format.

Leadership Experience

Upjao Agrotech

- **Prepared proposals and secured partnerships with Nvidia and Microsoft Azure Founders Hub**, resulting in **over $200,000 credits** in AWS and Azure cloud, enabling significant cost savings for the company.
- **Led and mentored a team** of over six engineers in the development & deployment of computer vision models in web application servers, and edge servers ensuring scalability and reliability.
- Designed new features with a focus on scalability and cost efficiency, **collaborating with cross-functional teams** (sales & testing) to ensure on-time, bug-free releases.
- **Spearheaded** the **development** and implementation **of an in-house Serverless Inference and Deployment Platform**, streamlining the management, version control, and deployment of diverse AI models (computer vision, NLP, multi-modal). This platform **enabled the creation of complex inference pipelines with parallel execution capabilities**, optimizing resource utilization, enhancing performance, and maintaining minimal latency during multiple inference calls.

- Part of the Startup core team, setup the initial infra and the designed the architecture from scratch.

Tata Consultancy Services

- Led a team of five engineers to **develop multiple proof-of-concept (POC) solutions from scratch.**
- Architected and implemented **cloud infrastructure**, including **high-performance computing (HPC)** systems, to support product development and deployment.
- **Communicated and collaborated effectively** with senior leadership and key stakeholders to present proof-of-concept (POC) solutions, highlighting their **viability, market value, and customer demand**.
- **Managed the full lifecycle of POCs**, from planning & design to development & execution, ensuring on-time, delivery within budget.

Academic Projects

Image Super Sampling | Python, PyTorch, OpenCV, CUDA

- Implemented a Deep Learning (DL) Model grounded in the **Super Resolution Generative Adversarial Network (SRGAN) research paper**, achieving notable success in upscaling a 720p movie to a high-resolution 4K format.
- **Optimized inference memory usage,** reducing GPU memory consumption by 40% compared to baseline.
- Enabled efficient processing of large videos without encountering OOM errors.
- Implemented **dynamic memory allocation** mechanisms to seamlessly transfer data **between CPU and GPU memory** during inference.

Age Prediction Server | Python, TensorFlow, OpenCV, Keras, Flask, CUDA

- Engineered a **ResNet50**-based model for discerning the age of a human face, leveraging the **Wiki CelebFace dataset** to train the system effectively.
- Executed meticulous **data cleaning procedures** to eliminate undesirable and unreliable data, **enhancing the dataset's quality for robust model training**.
- Established a **Flask Server** and utilized **NVIDIA's Triton Inference Server** for deployment. Conducted **stress testing on a single GPU node** to ensure efficient handling of parallel requests.

Real-Time Noise Cancellation in Audio Signals using Deep Learning | Python, PyTorch, Torchaudio, Librosa

- Utilized Mozilla's Common Voice Dataset for clean speech and the UrbanSound8k dataset as real-life noise to develop an algorithm employing Short-Time Fourier Transform (STFT) for intelligent mixing of the two, resulting in a noise-added speech signal.
- Implemented a **Wave-UNet-based DL model**, incorporating **One Cycle Learning Rate** (LR) to expedite training and achieve a global minimum faster. The trained model demonstrated impressive results, yielding an **MSE loss of approximately 0.0002**, while the **noise-added signal exhibited an MSE close to 0.1 compared to the noiseless signal.**

Drone (Quadcopter) | Pixhawk, BLDCs, F450 drone frame, 40 Amp ESCs, Telemetry device, 6200 mAh battery, GPS & Compass Module, Altimeter

- Acquired an in-depth understanding of key components, such as **brushless motors and ESCs**, essential for enhancing the flight time and telemetry range of quadcopters.
- Utilized the **PixHawk open-source board**, renowned for **autonomous vehicles**, as a foundational framework to construct a highly capable **quadcopter with advanced functionalities**.
- Designed and constructed a **quadcopter** that could fly higher than a seven-story building, incorporating a **power distribution board**, **ESCs, brushless motors, and a PixHawk control board**.
- Successfully configured the PixHawk control board to achieve a **flying range of 2.8 kilometers**, constant **telemetry input to the ground control station**, and a flight time of 20 minutes (**6200mAh battery**) with a **payload capacity of 6 kilograms**.
- A few features of the drone: Altitude Hold, RTL (Return to Land).

Line Following Rover | Arduino microcontroller, Pololu QTR-8RC, Motor drivers, Motors.

- The **color detection sensor** (Pololu QTR-8RC) is used to detect the color of the line on the ground. The **Arduino microcontroller** reads the output of the color detection sensors and sends signals to the motor drivers to control the motors. The motor drivers then control the motors to move the rover along the color.

Awards and Achievements

- Awarded **Best Team Award** in the TCS Rapid Labs Cohort, demonstrating exceptional technical and collaborative skills.
- Invited to serve as a **juror on the 2023 LaElevitia Hackathon Event**, attended by over 80 students, showcasing technical expertise and leadership potential.
- **Core member** of the University's **Innovation Centre**'s LaElevitia festival, hosting multiple events and demonstrating organizational skills.

Patents

- **System for Quality Assessment of Agricultural Product(436541)**
- **A method and system for encoding and decoding the data by using markers (420903)**
- **An Adjustable Apparatus for Uniform Image Acquisition. (Provisional - 202221022993)**
- **Generic Classification using Very Less Data (filed)**
- **Ear Tag Recognition System for the Cattle Identification. (filed)**
- **Cattle Identification using Face Recognition. (filed)**

Certifications

- Deep Learning Specialization, Convolutional Neural Networks, Sequence Models, Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization, Structuring Machine Learning Projects