Saket Karve

Bay Area, CA | +1-(215)-500-3499 |saketk1.sk@gmail.com | Technical Publications

SUMMARY

Highly skilled **Software Engineer** with 6+ years of experience in **Robotics** and **Machine Learning**, specializing in bridging the gap between advanced research and production-grade engineering. Proven track record in architecting **large-scale data infrastructure** and high-performance **inference systems** for foundation models like **VLM/VAMs**. Expert in building robust **Airflow** pipelines, optimizing **GPU utilization** for training, and deploying low-latency solutions on-device. Strong academic background in **Deep Learning** and **NLU**, with a focus on creating scalable, end-to-end systems for complex robotic action planning and perception.

PROFESSIONAL EXPERIENCE

Member of Technical Staff, Engineering | Dyna Robotics, Redwood City **August 2025 - Present**

- **Hybrid Data Ingestion & Scale:** Architected a multi-modal pipeline on **Airflow** supporting low-latency stream processing for real-time in-house data alongside a **high-throughput MapReduce framework**; designed and developed an optimized batch processing pipeline to scale processing of internal as well as externally acquired data at a throughput of **3,000 video hours per hour**, preparing massive datasets for VLM/VAM foundation model training.
- **Distributed Data Loading & Training Optimization:** Engineered a cross-cluster data loader and caching system to **minimize GPU data-access latency** during large-scale training jobs; developed an intelligent scheduling system to dynamically route workloads based on cluster utilization, dataset location, and GPU availability, drastically reducing training bottlenecks while maintaining granular observability over infrastructure health.
- **High-Performance Inference Infrastructure:** Developed specialized on-robot inference servers utilizing **TensorRT** for memory-efficient, low-latency trajectory execution, alongside a parallel raw PyTorch path to bypass conversion-related performance drops; integrated a targeted validation framework to secure data integrity from robot sensors to final episodes.
- **Tele-operation & Collection Stack:** Streamlined in-house data collection workflows by re-engineering the robot interface and optimizing episode file compression, resulting in significantly faster data uploads and increased throughput of high-fidelity datasets.

Software Development Engineer | Amazon Astro, Amazon Lab126, Sunnyvale **June 2020 - July 2025**

- **Robot Memory and Summarization:** Designed and implemented a framework to generate textual summaries of robot observations and interactions from up to 30 days, creating condensed representations for use in robot action planning leading the project with a team of 5.
- **Summary Evaluation:** Designed and built a dataset for evaluation of the generated summaries to compute various metrics like data loss, correctness, etc. Developed techniques that utilize a combination of classical methods and LLM based methods to compute metrics.
- **Robot Action Planning:** Designed and Developed end to end algorithms for robot action planning and world modeling using Large Language Models as part of an early stage prototype team responsible to build the architecture of LLM integration with the robot's brain.
- **Floor Plan Generation:** Contributed to the Floor Plan Algorithms team, developing algorithms to generate floor plans from SLAM data using geometry and computer vision and deploying the algorithm service to production that improved floor plan quality for 8000 devices.
- **Spatial Query Engine:** Architected a high-performance spatial query engine and a custom domain-specific language (DSL) to optimize the execution of complex floor plan queries; enabled enterprise clients to programmatically retrieve spatial elements—such as zones, poses, walls, and spaces—and overlay sophisticated business logic to automate plan generation.
- **Find Person Service:** Implemented an algorithm to generate an optimal path for the robot to find a person within the home's floor plan.

Software Development Engineer Intern | Amazon Halo, Amazon Lab126, Sunnyvale **May 2019 - Aug 2019**

- **Data Annotation Pipeline**: Developed a Flask-based web tool for data pipeline automation to aid training computer vision models for Amazon Halo. The web tool was deployed for engineers and scientists developing the models to upload data, for annotators to annotate the data and feed the annotated data into the data pipeline eventually used for training vision models for health tracking.

Research Intern | Indian Institute of Technology, Bombay **May 2017 - June 2017**

- Developed a **crowdsourcing platform** for the National Virtual Library of India (NVLI) project using Flask, implementing REST APIs to integrate open-source services like PyBossa to help analyze user responses to personalize the platform experience with ML algorithms.

PUBLICATIONS

Conceptor Debiasing of Word Representations Evaluated on WEAT **GeBNLP ACL 2019**

- Developed an effective technique using conceptors to debias pre-trained word embeddings with respect to typical human stereotypes like gender or race.
- Tested the debiased word representations for bias quantified using WEAT score. Demonstrated the utility of such representations on various NLP tasks.

Semantic Relatedness Measurement Using Modified Normalized Google Distance **DAL 2018**

- Developed a novel semantic relatedness method by modifying Normalized Google Distance (NGD) to integrate WordNet's Brown Corpus information content with Wikipedia occurrence statistics.
- This approach effectively quantifies relatedness for non-dictionary terms like jargons and proper nouns, outperforming the PMI measure by providing a more unbiased and normalized score.
- Experimental testing against human intuition confirmed the model significantly correlates with human judgment across diverse word pairs.

Comparative study of feature extraction techniques for face Recognition **ICCICT 2018**

- Designed a face recognition system comparing statistical feature extraction methods including Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Factor Analysis (FA) using four distinct classifiers.
- Evaluated performance on the ORL dataset, finding that Factor Analysis outperformed other techniques by achieving up to 100% accuracy with Neural Network, SVM, and Naive Bayes classifiers. Furthermore, established that combining PCA and ICA components significantly improved recognition sensitivity to facial expressions and orientations compared to using them independently.

A DSP based reprogrammable architecture for standalone signal processing applications **ICNTE 2017**

- Developed a standalone, reprogrammable hardware emulator for signal processing that eliminates host-PC dependency by executing user-defined algorithms directly on a TMS320F28069 DSP.
- The approach utilizes a custom diagramming interface to generate a structured Netlist, which is then decoded by the onboard system to emulate complex signal flows with a time complexity of O(n).
- Experimental results validated the architecture's efficiency, achieving an average computation time of 80 µs and supporting a signal bandwidth from DC to 1.25 kHz.

ACADEMIC PROJECTS

Retrospective reading using Dynamic Memory Networks for Question Answering | University of Pennsylvania **Mar 2020**

Developed a deep learning model for reading comprehension inspired by human reading strategies, featuring a **skimming retrospective module** and a **verification module for unanswerable questions**. The model uses **explicit reasoning via the Dynamic Memory Network** and was evaluated on the Microsoft NewsQA dataset.

Point Cloud Classification using using Graph Convolution Networks | University of Pennsylvania **Sept 2019**

Developed a PyTorch-like deep learning pipeline in CUDA from scratch. Built and trained a graph convolution network for 3D point cloud classification on ModelNet, achieving approximately **79% accuracy on test data** and a **40x GPU speedup**.

PennCloud | University of Pennsylvania **Dec 2018**

Developed a **fully distributed, replicated, fault tolerant** and **multi-threaded email** and **storage server** with an interactive user interface. The system allowed users to send and receive mail, upload and retrieve any type of file organized in a hierarchical manner.

SKILLS

- **Programming:** Python, C++, Java, CUDA, Matlab, Android, GPU Programming
- **Technologies:** PyTorch, AWS, GCP, Airflow, TensorRT, VLA, VAM, Alluxio, OpenCV, Large Language Models, RAG, Vector Databases, Knowledge Graphs
- **Area of Interest:** Robotics, Machine Learning, Computer Vision

EDUCATION

Master of Science and Engineering, Computer and Information Science **Aug 2018 - May 2020**

University of Pennsylvania Philadelphia, PA

Bachelor of Technology, Information Technology **July 2014 - May 2018**

Veermata Jijabai Technological Institute Mumbai, India