VARUN SREENIVASAN

varunsreenivasan16.github.io/vs/ | linkedin.com/in/varun-sreenivasan/ | (608) 422-2785 | vsreenivasanofficial@gmail.com

EDUCATION

M.S. Computer Science, University of Wisconsin-Madison *(Dec 2021)*
G.P.A: ***4.00/4.00***

B.S. Computer Science (Minor: Mathematics), University of Wisconsin-Madison *(May 2020)*
G.P.A: ***3.90/4.00*** *"Distinction in the Major"*, and *Dean's List*

LICENSES & CERTIFICATIONS

Certified Solutions Architect Associate, Certified Machine Learning Speciality, Certified Cloud Practitioner – Amazon Web Services

RELEVANT COURSEWORK

Artificial Intelligence, Machine Learning, Computer Vision, Data Science, Computer Networks, Mobile & Wireless Networks, Discrete Mathematics, Cryptography, Combinatorics, Linear Algebra & Differential Equations, Operating Systems, Algorithms, Data Structures.

EXPERIENCE

DataChat, Inc. **Madison, WI**
Senior Software Engineer I - Release Team **Jan 2025 - Present**

- Proposed and implemented the adoption of the Artillery load testing tool to evaluate the performance and scalability of DataChat's SaaS application. The testing helped uncover system bottlenecks that impeded application performance under load.
- Spearheaded the push towards CI/CD to streamline the release management process involving upgrades of Kubernetes hosted environments. This effort helped increase upgrade cadence and reduce manual effort.
- Led a team of developers to overhaul DataChat's Vector Database library and services, enabling support for VertexAI Vector Search and Milvus, which were critical providers to meet business demands.

Software Engineer II - Machine Learning (ML) & Generative AI (GenAI) Teams **Feb 2022 - Dec 2024**

- Spearheaded design and implementation of backend solutions to expand product capabilities within the ML & GenAI space.
- Enhanced the scalability and efficiency of the GenAI platform by adding Kubernetes' Horizontal Pod Autoscaling (HPA) for all GenAI services and optimizing CPU and RAM usage. The improvements led to a sizable reduction in Kubernetes related cloud costs.
- Accelerated LLM evaluation by setting up an LLM evaluation framework, enabling faster adoption of LLMs across various use cases.
- Increased GenAI availability by setting up the Azure Application Gateway to pool OpenAI's regional LLM resource quotas.
- Prevented the possibility of DoS attacks and rate limit errors via a ticketing service to control access to limited LLM resources.
- Led the research into the feasibility of hosting DataChat's custom LLMs across various cloud providers and platforms.
- Implemented the AutoML suite that allows users to train and evaluate multiple ML models and find the best one for their dataset.

National Science Foundation – IRIS HEP **Berkeley, CA**
Fellow – Graph Methods for Particle Tracking (High Luminosity Large Hadron Collider) **May 2021 - August 2021**

- Performed feature engineering in the processing stage to select cluster features to construct events from TrackML dataset.
- Developed an embedding pipeline to find a distance metric between 3D hit measurements pairs, optimizing particle differentiation.
- Accelerated the nearest neighbors search by replacing Facebook's Faiss with Fixed Radius Nearest Neighbors (FRNN) on CUDA.
- Optimized a PyTorch-based embedding model with 99% efficiency and 1% purity through hyperparameter scanning.

Citrine Informatics **Redwood City, CA**
NextGen-Fellow – Computational Materials Science **May 2018 - August 2018**

- Multi-university research project: Competitively selected, successfully completed bootcamp & workshop at Stanford University.
- Employed a Keras implementation of RetinaNet object detection model to identify defects in metals.
- Developed evaluation pipeline to determine recall and precision metrics. Obtained a model with 85% precision and 68% recall on the test set through hyperparameter optimization. Performed analysis to determine the reason for high false negative rate.
- Presented the results at NextGen Research Symposium in Golden, CO and co-authored a paper.

SOFTWARE PROJECTS

Autonomous RC Car
Helped develop a proof of concept for an open source, affordable, and performant autonomous vehicle testbed by training a PyTorch based SSD-Mobilenet object detection model on a custom dataset to do live detection of traffic signs such as speed limits and stop signs. The co-authored paper was accepted and presented at the ACM MobiArch 2022 Conference.

Business Success/Viability Forecast
Developed multiple ML models (Logistic Regression, Random Forest, KNN, Naïve Bayes, SVM, and Neural Net) using the Yelp dataset to predict whether businesses will survive the impacts of COVID-19. Performed feature engineering to obtain a final parsed dataset, created a training and validation pipeline that integrates SMOTE (address class imbalance) to guide parameter selection, evaluated models on unseen test data with multiple metrics, and determined the vital features using the Permutation Importance algorithm.

LANGUAGES & TECHNOLOGY SKILLSETS

Python, Java, Git, Docker, Kubernetes, AWS, GCP, Azure, CircleCI, Terraform, Helm, VS Code, Jupyter Notebook, Unix, MacOS