Vamshi Nagireddy

+1 (916) 707-2957 ◇ vamshi.knagireddy@gmail.com ◇ linkedin.com/in/vamshinr ◇ San Jose, CA

SUMMARY

- machine learning Engineer with extensive experience at Intel, designing benchmarking framework and driving Optimization of GenAI/non-AI workloads using Python and C++, leading to performance gains across Intel's hardware portfolio.
- Proven track record in improving the accuracy of resume parsing tool by 10% through optimizing advanced sequence models.
- Skilled in building job recommendation engines and integrating Elasticsearch for enhanced job matching and ranking, and deploying robust data pipelines with Kafka and MLFlow.
- Strong foundation in Python programming, deep learning models, and scalable architectures.
- Delivered robust solutions for automation across platforms playing a key role in UI development for benchmark applications.
- Proven expertise in MLOps, cloud technologies, and cross-functional collaboration to deliver scalable, data-driven solutions with measurable business impact.

SKILLS

- **Programming Languages** - Python, C, C++, Java, Shell Scripting.
- **Web Development** - HTML/CSS, JavaScript, Angular, React, Flask, Django, FastAPI, windows, Linux, Macos.
- **Machine Learning And AI** - TensorFlow, PyTorch, scikit-learn, Pandas, NumPy, ONNX.
- **DevOps And Cloud Technologies** - Docker, Kubernetes, Jenkins, Git, Kafka, MLFlow.
- **Data Analytics And Visualization** - Tableau, SQL, NoSQL (MongoDB, Neo4j, DynamoDB, Cassandra).
- **Build And Automation Tools** - CMake, Shell Scripting.

EXPERIENCE

Software Engineer — Intel Corporation — San Jose, CA **July 2023 - Present**

- Developed AI/non-AI workloads using Python and C++ for various consortiums to develop benchmarks and industry standards contributing to **performance gains** across **intel's hardware platforms**, including **CPUs, GPUs, and NPUs**.
- Architected and Implemented **event driven state machine framework** in Python for automating workloads across various platforms contributing extensively to Python **multi-threading** to manage concurrent processes (**battery logging**, running **workloads**) along with building a **HWINFO parser tool** to identify and address optimization areas.
- Contributed to Creating a **cross-platform security and installer tool** in C++ by manipulating **SQL** database and employing **CMake** for build automation. Collaborated cross-functionally with research scientists, hardware engineers, and software developers to align **AI workload strategies** with Intel's product roadmap.
- Evaluated Generative AI approaches in LLM inferencing for system integration, emphasizing limitations and advantages.
- Experimented with various **RAG pipelines, prompt engineering, model quantization, and fine-tuning** to optimize performance and adaptability contributing to **Optimization** of **Intel's AI software stack (OpenVINO) for LLM inferencing.**

Machine Learning Intern — Intel Corporation — San Jose, CA **May 2022 - August 2022**

- Conducted **inference optimization** using **ONNX Runtime** and **OpenVINO** for SPEC benchmarks.
- Developed and optimized ML workloads using **AVX512** instructions and **hardware accelerators**, including **face recognition**. Built a **performance parser** to compare workload efficiencies across platforms

Graduate Research Assistant — CSUS — Sacramento, CA **January 2022 - May 2023**

- Developed **data scraping** solutions and performed **data analysis** on SEC filings using Python.

Machine Learning Engineer — Phenom — Ambler, PA **July 2019 - July 2021**

- Spearheaded the Design and optimization of sequence models, including **RNNs, CNNs, LSTMs**, and **Transformers**, to improve Named Entity Recognition (**NER**) and text classification for a resume parsing tool, achieving a 10% accuracy increase.
- Developed **job recommendation engine** with **Elasticsearch** DB integration to improve job matching and ranking algorithms.
- Designed a **BERT** based sequence classifier for distinguishing various sections of a resume along with extracting raw text from unstructured data Building a data pipeline integrating parsed log data with **Kafka** and visualized insights using **ELK Kibana**.
- Utilized **MLFlow** for model training, Monitor the performance of deployed models, maintain, and update as necessary.

PROJECTS

- **Job Recommendation Engine With NEO4J, MongoDB, DYNAMODB, CASSANDRA.** Conducted a comprehensive study on NoSQL databases to evaluate their performance for job recommendation use cases. Generated insights into trade-offs between query execution times, replication strategies, consistency, availability, and performance in distributed NoSQL systems proposing recommendations for database selection based on application-specific performance requirements. (Weblink)
- **Human Activity Recognition Using Attention.** Implemented an attention-based BiLSTM-CNN ensemble algorithm to identify human gestures for monitoring health by Preprocessing raw CSI data captured from accelerometer, gyroscopes.(Github)
- **NLP on News Articles.** Extracted thousands of news articles through web scraping with bs4. Performed extensive data cleaning, exploratory data analysis on news articles data. Conducted text summarization with Spacy, text complexity analysis, and sentiment analysis. (Github)

EDUCATION

Master of Computer Science, California State University Sacramento 2021 - 2023

Relevant Coursework: DS Algorithms, Machine Learning/NLP, Distributed Systems, Computer Vision; 3.7 GPA