Rakesh Nagaraju

Santa Clara, CA | +1 (669) 288-4508 | rakesh.nju2205@gmail.com | linkedin.com/in/Rakesh-Nagaraju|
github.com/Rakesh-Nagaraju | http://www.rakesh-ai.com

SUMMARY

Engineer with over 5 years of experience in AI research and software development, specializing in computer vision, generative AI, NLP, and LLMs. Proven ability to deliver innovative solutions, lead teams, and optimize performance.

EDUCATION

MS in Computer Science, San Jose State University, San Jose, CA **Aug 2019 - May 2021**

- **Coursework:** Machine Learning, Artificial Intelligence, Distributed Computing, Cybersecurity

PUBLICATION

Published a paper titled 'Generating Fake Malware Using Auxiliary-Classifier GAN for Malware Analysis' (arXiv:2107.01620, 2021).

TECHNICAL SKILLS

- **Expertise:** Computer vision | Generative AI | NLP | LLM | Multimodal AI
- **Coding:** Python, C/C++, JavaScript, Shell Scripting, HTML/CSS, Fast API
- **Frameworks & libraries:** TensorFlow, PyTorch, PyTorch Lightning, HuggingFace, FastAPI, NumPy, Pandas, SciPy.
- **Tools:** AWS (EC2, S3, SageMaker, Lambda), GitLab CI/CD, MLflow, Docker.

EXPERIENCE

AI Design Engineer - Research – Uniquify Inc - Santa Clara, USA **Aug 2021 – Present**

RAG based Chatbot for internal SoC documents - *MinIO, Milvus, LlamaIndex, Chainlit, Docker*

- Developed and led the implementation of a RAG (Retrieval-Augmented Generation) chatbot for querying and interacting with Internal SoC documents, utilizing both dense and sparse retrieval methods to improve accuracy and response speed.
- Fine-tuned LlamaIndex for efficient document indexing and retrieval, while leveraging MinIO for secure document storage and Milvus for high-performance document similarity search.
- Deployed the system with Docker for scalability and ease of deployment, and integrated a real-time Chainlit front-end to enhance SoC team efficiency with dynamic, contextual responses to document queries.

Real-time face recognition and analysis on cameras - *Retinaface, Deepface, VGGFace, MLFlow*

- Successfully led a team in developing and productionizing face detection and face analysis models for low-latency, and high performance on CCTV cameras.
- Fine-tuned analysis models (age, gender, race, emotion) based on VGGFace on custom data, resulting in a 15% increase in accuracy.
- Optimize Retinaface for runtime performance from 2 images/s to 5 images/s by batching inference and quantizing weights.
- Streamlined the process of deploying and managing machine learning models by hosting inference on MLflow, ensuring efficient and scalable model deployment.
- Experience in training, fine-tuning and deploying SOTA models like SAM, ViT, CLIP, and multimodal such as mPLUG, GIT, BLIP for various applications including image captioning, training interns, and upskilling.

Person and object detection on SOC- *YOLOv3, YOLOv8, CNN*

- Successfully implemented training of YOLOv3 as in paper achieving ~35% mAP on COCO dataset.
- Improved detection and runtime capabilities by migrating current pipelines to run on YOLOv8 models.
- Achieve seamless real-time performance on embedded devices without GPU by using methods like quantization, freeze-training.
- Implemented pipeline on AWS using S3, EC2 for hosting on cloud..

Comprehensive AI Training Curriculum Covering Advanced AI concepts *– CV, LLM, NLP, CV, machine learning, deep learning*

- Developed extensive training materials, including over 70 Jupyter notebooks, encompassing detailed and advanced concepts in computer vision, NLP, machine learning, deep learning, LLMs, and MLOps.
- Created hands-on modules for tasks such as text summarization, question answering, named entity recognition (NER), text classification, modeling, training, fine-tuning, evaluation, and scalable deployment across a wide range of datasets.
- Provided thorough feedback and conducted evaluations to ensure interns' mastery of the material, effectively preparing them for successful transitions into industry roles as proficient AI engineers.

Defect detection system - *YOLOV4_Tiny, pytest, GitLab CI/CD*

- Finetune YOLOv4_Tiny model to detect defects in chip manufacturing with 98% accuracy, reducing chip testing cost by 50%.
- Collaborated with cross-functional teams to identify and resolve defects, ensuring deployment met the highest standards.
- Automated test pipelines, wrote and ran regression tests on models using pytest, CI/CD pipeline.

Senior Software Engineer – Capgemini - Bengaluru, India **Nov 2016 – Aug 2019**

Backend software development and maintenance - *python, DB2, beautifulsoup, OOP*

- Developing, updating, and maintaining customized Python software applications to extract tax-related data from a DB2 database.
- Performing operations on the data to align with business needs, creating APIs for access, and deploying them.
- Supporting User Acceptance Testing (UAT) for successful deployment of software versions.