VARUN REDDY CHANDA

AI/ML ENGINEER

+ 1 (863) 888-2030 | vchanda1006@gmail.com | Portfolio

PROFESSIONAL SUMMARY

AI/ML Engineer with expertise in **deep learning**, **computer vision**, and **NLP**. Adept at designing and optimizing **scalable AI systems**, with hands-on experience designing, training, and deploying machine learning models for **fraud detection**, **model compression**, and **Computer Vision** applications. Proficient in **Python**, **PyTorch**, **TensorFlow**, **SQL** and **Cloud technologies (AWS, Git).** Passionate about building **high-performance ML models**, improving computational efficiency, and deploying AI-driven solutions in **real-world applications**. Strong analytical thinker with excellent problem-solving skills, capable of breaking down complex challenges into actionable solutions. Skilled in collaborating with cross-functional teams, applying data science and engineering expertise to develop cutting-edge, data-driven solutions.

TECHNICAL SKILLS

- **Languages:** C, C++, Java, Python, JavaScript, SQL, HTML/CSS.
- **ML & Deep Learning:** PyTorch, TensorFlow, Scikit learn, Numpy, Pandas, Matplotlib, Seaborn, Ollama.
- **Cloud & DevOps:** AWS (EC2, S3, Lambda), Git, GitHub, CI/CD, Agile Methodologies.
- **Databases:** NoSQL (MongoDB), MySQL, SQLite, Pinecone (Vector DB).
- **Web & APIs:** Flask, React.js, Streamlit, Node.js, Express.js.

WORK EXPERIENCE

ML Engineer

TCR Innovations – Navi Mumbai, IN **June 2021 – Dec 2021**

Responsibilities:

- Developed scalable and interpretable ML models, optimizing efficiency for fraud detection, achieving **99.96%** accuracy in classifying fraudulent transactions.
- Designed and optimized multiple models, including **Logistic Regression (99.90%**), **Decision Tree (99.96%),**
- **Random Forest (99.96%),** and **Gaussian Naive Bayes (99.19%).**
- Analyzed **6 million+** transactions using **NumPy** and **Pandas**, reducing computation time by **30%**.
- Collaborated using **Git** and **GitHub** for version control, ensuring seamless integration, code review, and efficient deployment of ML models in a production environment.
- Demonstrated strong communication skills by effectively presenting model performance, **CI/CD** deployment updates, and technical challenges in **Agile** stand-ups.

Environment: Python, Numpy, Pandas, Tensorflow, Pytorch, GitHub, CI/CD.

PROJECTS

RAG System Using DeepSeek

- Developed a hybrid Retrieval-Augmented Generation **(RAG) system** integrating **FAISS** (vector search), **BM25** (sparse retrieval), and **Knowledge Graphs**, improving document retrieval relevance by **40%.**
- **Graphs**, improving document retrieval relevance by **40%** compared to standard keyword search.
- Leveraged **DeepSeek LLM** for hypothetical document generation, increasing relevant document retrieval by **25%,** leading to more contextually aware search results.
- Implemented voice-based query support using **WhisperX**, enabling real-time speech-to-text transcription for seamless interaction.
- Automated document processing pipeline using **PyPDFLoader** and **TextLoader**, enabling seamless extraction, Resulting in reduced manual text extraction efforts and improved data ingestion efficiency.

- Designed an interactive **Streamlit UI**, allowing users to upload documents, enter text/voice queries, and receive **AI-generated** contextual responses.

Environment: Python, PyTorch, WhisperX, Streamlit , Ollama (DeepSeek), LangChain, Pinecone

Filter Pruning in Deep Neural Network

- Reduced computational costs by **63%** by pruning redundant filters in the **VGG-16** model using **Hierarchical Agglomerative Clustering (HAC)**.
- Achieved **91%** test accuracy on **CIFAR-10** while pruning **63%** of the filters, with only a **2% accuracy** drop compared to the baseline model.
- Evaluated the impact of activation functions (**ReLU, Leaky ReLU, Tanh**) on model performance, achieving optimal trade-off between accuracy **(93.26%)** and pruning efficiency (**28.4%** filters removed).
- Enhanced model efficiency by reducing **FLOPs**, leading to a **40%** decrease in inference time, while maintaining high accuracy through **adaptive threshold-based pruning** of convolutional layers.

Environment: Python, PyTorch, Tensorflow, NumPy, Pandas, Matplotlib, VGG-16, CIFAR 10

Speed Bump Detection

- Developed a real-time object detection system using **YOLOv3** for autonomous vehicles, achieving **98% accuracy** with **TensorFlow** and **OpenCV.**
- Achieved **95% detection accuracy** using **SSD MobileNet V2** for identifying both marked and unmarked speed bumps in real-world road conditions.
- Enhanced road safety through **AI-driven** solutions, enabling real-time speed bump detection in less than **100ms per frame**.
- Collected and labeled a dataset of **50 images** using **Roboflow**, and **LabelImg**, improving annotation efficiency, model generalization, accuracy, and scalability.

Environment: Python, TensorFlow, OpenCV, Roboflow, YOLOv3, MobileNetv2

Vehicle Counting and Classification

- Achieved **98% detection accuracy** using the **YOLOv3** object detection model, ensuring precise vehicle identification, classification, tracking, validation, and segmentation.
- Implemented a **tracking algorithm** that maintained vehicle identities across frames with an **ID** persistence rate of **95%,** ensuring accurate vehicle counting, re-identification, and movement prediction.
- **Classified** vehicles into **5 categories**: cars, trucks, buses, motorcycles, and bicycles, providing detailed traffic composition, density, pattern, trend, and volume analysis.

Environment: Python, Tensorflow, Matplotlib, seaborn, YOLOv3, DeepSORT, OpenCV

EDUCATION

- Master’s degree in computer science from **Texas Tech University**.
    - GPA: 3.91

PROFESSIONAL PROFILES

- **Portfolio:** varunchanda.vercel.app
- **Linkedin:** www.linkedin.com/in/vchanda
- **GitHub:** www.github.com/varunreddy-ch