BHUMI GODIWALA

+1 9848109668 | godiwala.bhumi@gmail.com | https://www.linkedin.com/in/bhumigodiwala | https://bhumigodiwala.github.io

EDUCATION

University of Southern California — Los Angeles, CA
Master of Science — August 2021-May 2023
- Master of Science in Computer Engineering (Machine Learning and Data Science specialization) (GPA 3.8/4.0)

Dwarkadas J. Sanghvi College of Engineering, University of Mumbai — Mumbai, India
Bachelor of Engineering — July 2016-October 2020
- Bachelor of Engineering in Electronics and Telecommunications Engineering (CGPA 9.19/10)

TECHNICAL SKILLS

- Programming Languages: Python, Java, Git, C, C++, MySQL, SQL, Oracle, Object Oriented Programming (OOP/OOPs)
- Software: PyTorch, Jupyter Notebooks, Anaconda, PyCharm, Eclipse, Docker, JIRA, Node JS, AWS S3, AWS CLI, REST API, Spark
- Libraries and Frameworks: Matplotlib, Tensorflow, Keras, Numpy, OpenCV, Scikit-Learn, Pandas, Onnx, Hadoop
- Web-Technologies: HTML, CSS, Javascript, JSON

WORK EXPERIENCE

MemoryCare AI Inc — Newport Beach, CA
Senior AI/ML Engineer — May 2024-Present
- Developing custom LLM and RAG pipelines using Python and TensorFlow, enhancing cognitive assistance for TBI, Dementia, and Alzheimer's patients.
- Created an avatar-based platform with PyTorch and OpenCV to track speech distortions, behavioral changes, and tone, achieving 90% accuracy.
- Designed a MoCA test-inspired cognitive assessment platform on Microsoft Azure, reducing assessment time by 35% and ensuring scalability.

USC Information Sciences Institute — Marina Del Rey, CA
Machine Learning Engineer — July 2023-Present
- Developed an Adaptive Mixture Quantization (AMQ) scheme for Cloud/Edge AI systems, improving model accuracy by 5% and reducing model size by 15%.
- Demonstrated quantization strengths, balancing integer and float-point methods for compression and accuracy, leading to a 30% improvement in communication efficiency.

TetraMem Inc — Fremont, CA
Machine Learning Intern — January 2023-May 2023
- Created ML models (e.g., Visual Wake Words, Human Pose Estimation) with PyTorch, Onnx Runtime. Verified on a state-of-the-art AI inference chip with in-memory computing capabilities
- Designed a customized model with positive-only weights using PyTorch, maintaining accuracy. Applied Quantization Aware Training (QAT) for optimized models with reduced parameters

TetraMem Inc — Fremont, CA
Machine Learning Intern — May 2022-August 2022
- Built Human Pose Estimation ML models using PyCharm, Tensorflow, and COCO Python API in Linux environment. Validated on an in-memory computing AI inference chip post neural network optimization
- Executed post-training quantization to optimize developed Machine Learning models, achieving 93% accuracy with reduced size

Tata Consultancy Services - ION — Mumbai, India
Software Engineer — October 2020-August 2021
- Created JAVA, HTML, CSS, and JavaScript forms for university portals, deploying in real-time using TCS' framework
- Conducted metadata mapping, testing, reports generation, and optimized data segregation by course details

ACADEMIC PROJECTS

Automatic Email Generation
- Utilized FastAPI using GPT-3 for the backend and React at the frontend to automatically generate emails
- Deployed using Docker as well as AWS services like AWS Lambda, EC2 and API Gateway

ASL Gestures Prediction using ST-GAN for Shadow Removal
- Engineered pre-trained GAN-CNN fusion to enhance ASL Gesture classification, mitigating shadow effects
- Attained 92.9% test accuracy for 'E' and 'S' ASL Gestures classification. Utilized MLFlow for MLOps tracking and experience logging

Banking Subscription Analysis
- Executed client subscription prediction using supervised algorithms: Logistic Regression, Decision Trees, Random Forests, and SVM
- Assessed performance via Accuracy, F1 score, and confusion metrics. Evaluated semi-supervised techniques including S3VM, label propagation, label spreading, and Co-training classifier

Community Car Rental Platform
- Created car rental web app, deployed on Google Cloud CLI, and enhanced with Google Maps, Cloudinary, and VIN API integrations
- Utilized Hadoop for efficient storage and batch processing of large datasets, and leveraged Spark for real-time analytics to provide insights and recommendations based on car rental trends and vehicle availability. Designed database schemas via GraphQL with data hosted on MongoDB