Vijayraj Gohil

vrajgohil130@gmail.com | 989-403-2857 | www.linkedin.com/in/vijaygohil/

SKILLS

Programming Languages: Python, C++, CUDA, SQL, Bash
Technologies: PyTorch, JAX, HuggingFace (Transformers, Diffusers, PEFT), DeepSpeed, vLLM, TensorRT, ONNX Runtime, Docker, AWS (EC2, S3, ECR, SageMaker), Weights & Biases, PyTorch DDP, FSDP, Unsloth.
Libraries: OpenCV, NumPy,, scikit-learn, FlashAttention, LoRA/QLoRA, LangChain, Triton.

EXPERIENCE

Machine Learning Engineer Feb. 2023 – Present
Visual Concepts (2K Games) Kirkland, WA

- Led research on vision models for game-play animation synthesis and motion in-betweening, driving quality improvements across multiple animation segments.
- Designed and deployed an automated game processing pipeline — encompassing timer/clock detection, scripted data annotation, camera synchronization, and calibration — cutting total processing time by ~95%.
- Implemented inference optimizations using TensorRT, vLLM, and model quantization techniques, reducing latency by ~350% and enabling a ~281% increase in animation output for a major game release.
- Migrated the ML codebase from legacy Perforce to GitHub and re-architected the animation generation pipeline onto AWS (S3, EC2, ECR) with Docker, boosting operational efficiency by ~80%..

Machine Learning Engineer May 2022 – Dec. 2022
GRUBBRR Boca Raton, FL

- Implemented object detection models and few-shot learning for object recognition while reducing training data required for onboarding by ~96.5%.
- Refactored model evaluation code and experimented with various model quantization and sparsification for real-time inference using ApacheTVM Engine and TensorFlow Lite.
- Built state-of-the-art chat agent with Dialog Flow messenger widget using Language Modeling and Sentiment Analysis, providing customers with fast and accurate responses.

(**Analog AI) Research Assistant** Mar. 2022 – May 2022
IBM Research New York, NY

- Researched the Analog AI system to perform in-memory computing for faster training and inference.
- Benchmarked the in-memory computing library using PyTorch for different multi-GPU cluster settings with up to 16 GPUs on large-scale systems.

Machine Learning Researcher August. 2020 – Feb. 2021
Dwarkadas J Sanghvi College of Engineering Mumbai, India

- Classified apple leaves with Foliar disease by creating a multi-label classification model using Efficient-Net as the backbone and a custom set of fully connected layers.
- Optimized features and model selection by implementing feature engineering, data wrangling, and visualization.
- Experimented with different learning rate schedulers, early stopping, and other hyperparameters for faster convergence, which achieved ~87% test accuracy.

Machine Learning Software Development Intern May 2019 – Aug. 2019
Indian Institute of Technology Bombay Mumbai, India

- Developed image annotation application using MATLAB GUI for classification-based tasks.
- Designed a model for image classification using mixture of feature descriptors for different classes of vehicles.
- Created a method to transfer a trained model to a different platform and reduced manual intervention by optimizing the image formatting method, decreasing time by ~94% for ~20,000 small scale images.

PROJECTS

Bench-marking CIFAR-10 on a compressed CNN Model | Python, PyTorch, CNN, Multi-class Classification

- Constructed a machine learning pipeline for rapid experimentation for benchmarking with hyper-parameters and augmentation techniques, resulting in ~94.26% test accuracy and compressing the ResNet-18 to ~4.8M parameters.
- Trained a model on a distributed data center cluster in a multi-GPU setting for performance profiling with different training procedures.

Deep v-SLAM | Deep Learning, SLAM, Pose Estimation, Computer Vision

- Implemented a deep learning based visual SLAM algorithm for 3D camera pose estimation and map reconstruction.
- Utilized graph neural net (SuperGlue) feature descriptor for feature matching and made a custom integration with ORB slam for back-end reconstruction while utilizing bundle adjustment for refinement.
- Programmed the back-end pipeline from scratch for camera tracking and pose estimation via bundle adjustment

EDUCATION

Master of Science in Computer Engineering Sep. 2021 – Dec. 2022
New York University (NYU) New York, NY
Relevant Courses: Robot Perception, Robot Localization, Deep Learning, High Performance Machine Learning Systems, Probability
GPA: 3.7/4.0

Bachelor of Engineering in Electronics Engineering Aug. 2016 – Nov. 2020