JAIMIN MANIYAR

AI Software Solutions Engineer @ Intel

☎ +91 97266 71513
Bengaluru, Karnataka, India
✉ maniyar7jaimin@gmail.com

https://linkedin.com/in/maniyar2jaimin
https://github.com/maniyar2jaimin

Data Scientist with 7+ years managing the end-to-end AI solution lifecycle, applying ML, statistical analysis, and deep learning (LLMs, Transformers) using Python, PyTorch, TensorFlow. Increased LLM performance 2.5x on Gaudi hardware through targeted optimization. Led the development of scalable model evaluation tools adopted across 3 global sites, improving assessment efficiency. Created automation reducing data workflow effort by 80%. Applied advanced data analysis techniques to diagnose and resolve critical hardware/software flaws, enhancing system stability. M.Tech. in Computer Science (Data Analytics), VIT Vellore.

SKILL

- Python (NumPy, Pandas, Scikit-Learn, OpenCV)
- Machine Learning (SVM, XGBoost, Random Forest)
- Data Visualization (Tensorboard, Matplotlib)
- Large Language Models (LLMs)
- Deep Learning
- SQL
- Exploratory Data Analysis
- Statistical Modelling
- Data Analysis & Processing
- Generative AI (RAG, LangChain, vLLM)
- ETL
- PyTorch / Tensorflow
- Computer Vision
- Git / Gerrit / Jenkins
- Docker / Kubernetes(K8s)
- Natural Language Processing (NLP)

WORK EXPERIENCE

AI Software Solutions Engineer | Intel | Bengaluru

APR 2021 - PRESENT

- Increased Large Language Model (LLM) like Llama3 and GPT network performance by 2.5x on Intel Gaudi hardware (vs. competitor) through targeted PyTorch and hardware-level optimizations, enabling successful deployments for major CSP clients.
- Architected and led development (70% core code contribution) of a scalable model evaluation/optimization tool taken in at 3 global Intel sites, standardizing evaluation pipelines and improving assessment efficiency through automated data integrity checks.
- Resolved 11+ critical hardware/software interaction issues by conducting diagnostic analysis using system telemetry/model logs to identify root causes for LLM performance regressions.
- Engineered an ML-driven script for creation and configuration of containerized Kubernetes environments, accelerating data science workflow setup time by 40% (from 30 to 18 minutes).
- Accelerated deep learning model debugging cycles by 65% by integrating, Tensorboard plugins, PyTorch-Profiler and compiler-level graph visualizations, enabling rapid identification of performance bottlenecks.
- Conducted A/B testing on model optimization techniques, resulting in 7% improvement in model inference latency (without compromising accuracy), validated with profiling and visualization tools.
- Directed experiments optimizing model architecture, personally executing 5 different performance or accuracy trade-off studies, and implementing the highest ROI recommendation.

Deep Learning Software Engineer | Intel | Bengaluru

MAY 2018 - APR 2021

- Developed and deployed an automated defect screening system (OEM/ODM levels) using a PyTorch image classification model, meeting target Defects Per Million (DPM) rates of 2.5; co-authored a research paper on this system.

- Applied ML/DL techniques to production datasets, generating diagnostic data and utilizing the resulting insights to resolve flaws within neural network and hardware components (vMin, SRAM, MME), improving system stability by 95%.
- Authored custom Python kernels within TensorFlow/PyTorch frameworks, resolving critical hardware constraints and boosting neural network latency on AI inference hardware by 2x versus competitor's performance.

Graduate Intern Technical | Intel | Bengaluru — JUL 2017 - MAY 2018

- Researched and chipped in 90% to the TTVGAN project, resulting in a publication awarded Best Paper at the NCSET Conference in Nov 2017.

PROFESSIONAL ACHIEVEMENTS

Work Level Achievements

- Key Designer and course instructor for Level 1, 2, and 3 AI and Data Science courses in Intel.
- Led AI software stack validation as an SME for a top AI inference hardware I/P during critical bring-up phase in a high-impact, strategic on-site program in Haifa, Israel—contributing to successful system power-on.

Publications

- CS & IT Conference: “Screening Deep Learning Inference Accelerators at production lines” - Nov 2022.

Research Work

- TTVGAN: Teach Machines to be creative: Video Synthesis from Text using Generative Adversarial Network (Best Research Award - Dec 2016).
- Innovative Monitoring System for TeleICU Patients using Video Processing & Deep Learning (Best Research Award - Apr 2018).

Personal Projects

- Fish Image Classification: Enhanced real-time object detection for the Nature Conservancy fishery monitoring using YOLOv5, achieving a 15% improvement in mAP (mean Average Precision) on edge devices through model quantization.
- SMS Classification: Engineered AI-driven text classification pipeline utilizing NLP techniques, achieving a 95% accuracy rate in categorizing SMS messages during the Samsung-organized hackathon, demonstrating rapid model development skills.

EDUCATION

VIT (Vellore Institute of Technology), Vellore | M.Tech. in Computer Science | May 2018

- Specialization in Machine Learning and Data Analytics
- CGPA - 9.31 / 10

GTU (Gujarat Technological University), Gandhinagar | B.E. in Computer Engineering | May 2016

- CGPA - 8.46 / 10