Lakshmanan

Machine Learning Engineer

sivalakshmanan8@gmail.com
https://geekylax.github.io
9677695244
Chennai, Tamilnadu

Machine Learning Engineer with 5.3+ years in NLP, GenAI, and deploying LLMs using VLLM. Expert in Python, PyTorch, TensorFlow, and cloud platforms. Led a 15-member team, developing and optimizing models with advanced ML frameworks and strategic deployment strategies.

Core Skills

Programming Languages: Python, JavaScript, C++(Beginner),
Machine Learning & Deep Learning: TensorFlow, PyTorch, TensorFlow.js (TFJS), Scikit-Learn, NumPy, SciPy, Pandas,
Natural Language Processing (NLP): RoBERTa, GPT-4, BERT, vLLM,
API & Web Application Development: Flask, Django,
Cloud Platforms: AWS (SageMaker, Lambda, DynamoDB, S3, ECR, EKS, ECS, API Gateway, RDS), GCP (Vertex AI, Cloud Storage, BigQuery),
Data Management & Databases: MySQL, PostgreSQL, MongoDB, DocumentDB, Redshift, Redis,
MLOps & Deployment: Kubernetes, Triton Inference Server, Docker, CI/CD pipelines,VLLm,
Monitoring & Troubleshooting: AWS CloudWatch, Logging & Metrics

Education

Sriram Engineering College
Aug 2014 - Aug 2018

Bachelor of Engineering (ECE)
GPA 7.9

Karapettai Nagar Hr Sec School
Jun 2017 - May 2018

Higher Studies computer science
GPA 80

Languages

English *(fluent)*
Japanese *(read)*
Tamil *(fluent)*

Work Experience

Machine Leanring Engineer
Aug 2023 - Present

Numentica Technologies Pvt Ltd | *Bangalore*

As a AI/ML Engineer specializing in NLP and LLMs, I design and deploy scalable AI solutions, leveraging transformer models like RoBERTa, LLaMA, and GPT-4. Proficient in MLOps, fine-tuning, and optimizing models for efficient, production-ready deployment

- Developed and fine-tuned LLMs like RoBERTa and LLaMA for extreme multi-class text classification, achieving 94% accuracy. This significantly improved product categorization, enhanced branding, and boosted company revenue by optimizing the categorization of over 400,000 products.
- Implemented Retrieval-Augmented Generation (RAG) techniques, combining NLP with external knowledge bases to enhance model responses and improve retrieval accuracy in generative AI applications.
- Optimized transformer-based inference services in production, enhancing throughput by 14x through batching, parallel processing, and model quantization, ensuring scalable and efficient deployment.
- Streamlined MLOps practices, deploying models using TensorFlow and PyTorch on Triton Inference Server, ensuring robust model deployment, monitoring, and maintenance. Leveraged vLLM and CI/CD pipelines on GCP for optimized throughput and seamless integration into scalable AI applications.
- Designed and deployed GenAI applications using popular LLMs like OpenAI's ChatGPT, Gemini, and LLaMA, with expertise in prompting and fine-tuning for various use cases
- Trained LLM models using Google Cloud Platform (GCP) and AWS, leveraging services like Vertex AI, GCP's AI Platform, and AWS SageMaker. Utilized these platforms for scalable training, efficient resource management, and seamless deployment of models, ensuring optimal performance and cost-effectiveness

Machine Leanring Engineer (Lead)
Aug 2021 - Aug 2023

Anicca Data Science and Solutions | *Bangalore*

Led end-to-end AI/ML projects at Anicca Data, deploying scalable solutions on edge and cloud platforms, managing a 15-member team for computer vision tasks, and optimizing models for enterprise-

Certificates

Fast.ai Nov 2019

Coursera ML and DL Jan 2019
Coursera

Interests

MLOps and Model Deployment, Computer Vision, Deep Learning Research, Edge AI, Generative AI, NLP

scaledeployment using cutting-edge tools like TensorFlow Lite and PyTorch
Mobile

- Developed MVP applications showcasing AI/ML models, demonstrating their value to clients and stakeholders.
- Designed and engineered data pipelines and infrastructure using AWS and GCP services, such as S3, EC2, Cloud Storage, and BigQuery, to support scalable enterprise machine learning systems. This included automating data ingestion, transformation, and deployment processes for large-scale ML applications.
- Deployed ML applications on edge and cloud platforms (Intel Nucs, AWS, Azure), integrating CI/CD and automation
- Managed a 15-member team for a computer vision project, driving successful project delivery and innovation.
- Optimized and deployed models using tools like TensorFlow Lite, ONNX, and PyTorch Mobile for edge and mobile devices.
- Led the Vicco.app project for McDonald's, enhancing customer analytics through object detection, tracking, and scaling models to 200 stores

Machine Leanring Engineer
Jun 2019 - Aug 2021
Hubino Technologies | *Chennai*

Skilled in reviewing software code, managing AI datasets, solving problems with DL/ML, and delivering 20+ POCs. Expertise in Flask/Django, Agile, and SDLC tools like JIRA, Git, and TDD. Proven track record in customer-focused AI solutions

- Led the development of an AI pipeline that analyzed 10 million images, improving model efficiency and reducing inference time for camera health monitoring.
- Executed 20+ successful POC projects, demonstrating expertise in solving customer problems using deep learning and machine learning techniques.
- Implemented a real-time driver health monitoring system using wearable devices, enhancing safety by detecting driving infringements with deep learning models.
- Streamlined AI datasets creation and management, ensuring data quality and reliability for AI applications, contributing to improved project outcomes.
- Proficient in Flask/Django frameworks and Agile methodologies, with hands-on experience in JIRA, Git, Confluence, and Test-Driven Development (TDD) to support the software development lifecycle.

System Engineer
Jun 2018 - Feb 2019
DCKAP Technologies | *Chennai*

Developed ML pipelines on AWS, monitored and configured AWS and GCP servers. Experienced with AWS services (SageMaker, DynamoDB, Lambda, etc.) and infrastructure as code (Terraform, CloudFormation). Ensured optimal server configurations and seamless ML deployment.

- Developed ML Pipelines on AWS: Created and managed ML-based pipelines for robust model development and deployment.

- Server Monitoring and Configuration: Monitored AWS and GCP servers, ensuring proper configuration and performance.
- Proficient in AWS Services: Experienced with AWS services like SageMaker, DynamoDB, Lambda, and more.
- Expertise in Infrastructure as Code: Skilled in Terraform and CloudFormation for managing infrastructure.
- Optimized ML Deployment: Ensured seamless deployment and integration of machine learning models across cloud platforms.

Projects

Vicco.app - For McDonald's (Computer Vision Customer Analytics)

Present

Developed a customer behavior analysis system using CCTV footage, implementing object detection and tracking algorithms to monitor and analyze customer movements throughout the store

- High Accuracy: Achieved a highly accurate model through extensive annotation and analysis, enhancing business performance.
- Edge Deployment: Implemented quantization and pruning methods for model deployment on edge devices, optimizing performance and efficiency.

AVTS (Advanced Vehicle Tracking System)

Present

Developed an advanced vehicle tracking system for Anicca Data to provide real-time tracking and monitoring of vehicle fleets using machine learning and IoT technologies. The system integrates various data sources to enhance vehicle management and operational efficiency.

- Real-Time Tracking: Enabled real-time monitoring of vehicle fleets with high precision.
- Integrated Data Sources: Combined multiple data sources for comprehensive vehicle management and operational insights.

Order Accuracy for McDonald's

Present

Implemented a system to ensure order accuracy by preprocessing image data, identifying anomalies, and verifying order information to improve the ordering process

- Data Accuracy: Enhanced order accuracy by identifying and correcting discrepancies.
- Visual Insights: Used data visualization to present outcomes to senior leadership.

portfolio

Present

For More project Deatails Check here

- https://geekylax.github.io/