Thao Tran

MLOps | AI Engineer

+13502203350 thao1luon@gmail.com https://www.linkedin.com/in/thao-tran-54b148372/ Sejerslev, Denmark

SUMMARY

Machine Learning & MLOps Engineer with 8 years of experience in building and deploying advanced AI solutions across cloud environments.

- **Scalable ML Pipelines**: Specialize in designing, optimizing, and automating end-to-end workflows using **Python**, **Docker**, and **Kubernetes**.
- **Machine Learning Techniques**: Expertise in **NLP**, **Computer Vision**, and **RAG** systems.
- **AI Frameworks**: Hands-on experience with **TensorFlow**, **PyTorch**, and **AWS SageMaker**.
- **Cloud Solutions**: Proficient in **AWS**, **GCP**, and **Microsoft Azure**, with a focus on high-performance, cost-effective, and resilient architectures.
- **Model Deployment & Monitoring**: Expertise in **CI/CD automation**, **model lifecycle management**, and **cloud infrastructure** for seamless deployment and monitoring.
- **AWS Certified Solutions Architect - Professional**: Proven ability to design and implement efficient cloud architectures, ensuring scalable and secure machine learning deployments.

EXPERIENCE

MLOps Engineer

Remote

RapidScale

04/2023 - 04/2025

Project Name: AI-Driven Job Recommendation & Resume Parsing System

- Developed and deployed **scalable ML pipelines** using **AWS SageMaker**, **Docker**, and **Kubernetes**, ensuring smooth deployment of models for **job recommendation** and **resume parsing**.
- Built and optimized data pipelines using **AWS Lambda** for real-time data collection and **AWS Glue** for data transformation, automating the scraping of job postings and processing user resumes into a clean, usable format.
- Created and integrated **embedding models** using **Sentence-BERT** and **AWS SageMaker** to convert job descriptions and resumes into vector representations, enabling efficient **semantic similarity** matching for job recommendations.
- Implemented **CI/CD workflows** with **Jenkins** and **Terraform**, automating the deployment, versioning, and continuous integration of models, ensuring smooth updates and scalability of the machine learning solution.
- **Optimized inference pipelines** for job recommendations by using **TensorFlow Serving** and **Docker** containers, improving model response time and ensuring low-latency predictions in production.
- **Stored and managed embeddings** in **AWS S3** and **DynamoDB**, enabling efficient retrieval and storage of job and resume vectors for fast, real-time similarity matching.
- Monitored model performance with **AWS CloudWatch** and **SageMaker Model Monitor**, tracking changes in data and performance over time to trigger model retraining and ensure continued accuracy.
- **Built and exposed RESTful APIs** using **FastAPI** to serve embedding-based job recommendations, allowing seamless integration with backend systems for real-time processing.

ML Engineer

Remote

TikTok

04/2022 - 02/2023

- Spearheaded the development of an **AI-powered conversational assistant** to support creators and enhance in-app user experience across millions of global users.
- Customized and fine-tuned **Large Language Models (LLMs)** using TikTok-relevant data, optimizing for **multilingual understanding**, **content trends**, and **creator support**.
- Designed a **smart retrieval system (RAG)** to enable real-time responses by blending **vector similarity search** with **graph-based user interaction data** (Neo4j), improving contextual relevance by 30%.
- Built tools that could **understand slang, emojis, and cultural nuances**, using cutting-edge **NLP and NLU** techniques tailored to short-form content and viral language.
- Collaborated with content safety and moderation teams to implement **AI-assisted moderation flows**, ensuring accurate and scalable detection of policy-violating content.
- Leveraged **Azure AI and Databricks** to handle large-scale experimentation, model training, and deployment with near real-time feedback loops.
- Integrated **Azure OpenAI** services to explore generative capabilities like **automated comment summarization**, **caption suggestions**, and **AI-powered DMs**.
- Achieved a **20% latency reduction** for LLM inference by optimizing pipelines with **GPU acceleration**, ensuring smooth, real-time AI experiences within the app.
- Embedded the Gen AI system into TikTok's **creator tools**, boosting engagement and retention by providing instant, context-aware assistance and feedback.
- Regularly tested emerging Gen AI models and NLP strategies to ensure TikTok stays ahead in delivering **innovative, socially aware AI experiences** at scale.

EXPERIENCE

Data Scientist
Remote

Danske Bank
01/2021 - 03/2022

- Developed machine learning models for **credit risk assessment**, **fraud detection**, and **customer segmentation** using algorithms such as **Random Forest**, **XGBoost**, and **Neural Networks**.
- Preprocessed and transformed data using advanced techniques like **feature engineering**, **missing data imputation**, and **data normalization** to create high-quality datasets suitable for model training.
- **Implemented time-series forecasting models** for financial market predictions and customer behavior analysis using **ARIMA** and **LSTM** (Long Short-Term Memory) networks.
- Optimized model performance by applying **hyperparameter tuning**, **cross-validation**, and **grid search**, improving accuracy and reducing overfitting.
- **Automated end-to-end data pipelines** for data extraction, cleaning, and transformation using **Python** and **Apache Spark**, significantly reducing manual preprocessing time.
- Collaborated with **business analysts and financial teams**, ensuring that models aligned with business objectives and compliance regulations, and presented findings using **Tableau** and **Matplotlib** for visualization.

Computer Vision Engineer
Hong Kong

SenseTime
09/2019 - 12/2020

- Developed and optimized **CoreML-based computer vision models** for real-time inference on iOS devices, powering next-gen visual intelligence features in consumer-facing mobile apps.
- Contributed to the **SenseNova foundation model system** by integrating lightweight yet high-performance visual models into mobile environments, ensuring seamless user experiences at the edge.
- Leveraged the **SenseCore AI infrastructure** to streamline model training and deployment pipelines, reducing development time and operational cost for mobile CV solutions.
- Engineered and converted deep learning models (e.g., CNNs, object detection, segmentation) into **CoreML format**, optimizing them for on-device performance with **quantization, pruning, and model compression** techniques.
- Collaborated with iOS engineers to embed **real-time vision capabilities** — including face tracking, scene recognition, and AR overlays — into the company's flagship app.
- Designed custom preprocessing and postprocessing logic for the **CoreML models**, improving prediction accuracy and reducing latency by over 25% across iPhone and iPad devices.
- Helped enable industry use cases across **Smart Auto**, **AI-powered retail**, and **mobile AR**, aligning with the company's mission to build efficient, scalable AI across verticals.

Bachelor of Science in Computer Science
Beijing

Peking University
09/2015 - 07/2019

LANGUAGES

English (Native) Vietnamese (Native) Chinese (Proficient)

SKILLS

Amazon EC2, Amazon S3, AWS Lambda, AWS Redshift, AWS SageMaker, Azure, Azure Databricks, C/C++, CSS, Databricks, Deep Learning, Docker, ECommerce, GCP, Git, JavaScript, Jenkins, JIRA, Kafka, Kanban, Keras, MLOps, Neo4j, NLP, NLTK, OpenCV, Python, Tensorflow, PyTorch, CoreML, QuickSight, Scikit-Learn, Scrum, Sentiment Analysis, Tailwind