SALVATORE PETROLO

Machine Learning Engineer

toredev@outlook.it   +39 3887808493   t0re199.github.io   github.com/t0re199   linkedIn.com/t0re199

About

Innovative and results-driven Machine Learning Engineer with a Master's degree in Computer Engineering, specialising in Artificial Intelligence and Machine Learning. Two years of experience in a dynamic startup environment, leading the end-to-end development of advanced AI models and tools. Expertise in transformer architectures, natural language processing (NLP), Large Language Models (LLMs), RAG systems, computer vision, deployment of models on edge devices and on the cloud. Proven ability to optimise models for performance and resource efficiency.

Professional Experience

Machine Learning Engineer

February 2022 - Present

Knowlix GmbH - Tutzing, Bayern, Germany

Transformer Encoder Development: Led the full lifecycle development of a transformer encoder, handling data curation, selection, model training, and deployment. The pre trained encoder improved the downstream classification task accuracy by 10%.

Fine-Tuned LLM for Information Extraction: Supervised Fine-Tuning (SFT) of a large language model (LLM) to perform key information extraction from documents, improving data extraction accuracy.

Vision-Language Model Development: Developed a vision-language model (VLM) for complex document understanding tasks, integrating both visual and textual features.

Edge Deployment of Transformers: Created a small, efficient decoder-only transformer model, optimised for deployment on Apple mobile devices using CoreML. Implemented performance optimisations such as key-value caching and grouped query attention to enhance speed and reduce memory usage. Achieved 8x speed-up with respect to the standard implementation.

Federated Learning Pipeline: Designed and implemented a federated learning pipeline to continually improve models based on user feedback, ensuring privacy and security while enhancing model performance.

Document Segmentation AI: Built an AI model for segmenting documents into different sections, enhancing document processing capabilities.

Serverless REST Endpoint for LLM Interaction: Implemented a serverless RESTful API for interacting with a large language model (LLM) enabling scalable, real-time model access via cloud services.

Education

Master of Science in Artificial Intellingence and Machine Learning

University of Calabria - Cosenza, Calabria, Italy   September 2020 - July 2022

Thesis: Deep Anomaly Detection in ECG Signals to Detect Arrhythmias.

Score: 110/110 cum Laude

Bachelor of Science in Computer Engineering

University of Calabria - Cosenza, Calabria, Italy   September 2016 - September 2020

Thesis: Object Oriented Data Language: a language for developing dynamic data collection web application.

Score: 110/110 cum Laude

Acknowledgments

University of Calabria - Cosenza, Calabria, Italy   2022

Best Student Award for the 2nd Year Master's program at University of Calabria. [link]

Technical Skills

Programming Languages: Python, Java, Swift, Javascript, C, C++, CUDA, Assembly.

Model Architectures: Transformers, Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNS).

Machine Learning Libraries: NumPy, Pandas, SciPy, Scikit-Learn, SciPy, OpenCV, PIL, Matplotlib, Seaborn, TensorFlow, Keras, PyTorch, TorchVision, Transformers, Llama.cpp, Ollama, Vllm, Mlx, Apache TVM, MLC, JAX, ONNX, LLama Index, LangChain, Weights and Biases.

LLMs Optimisations: Keys-Values Caching, Grouped Query Attention, Quantisation, Speculative Decoding, Prompt Engineering.

Tools and Platforms: Git, Github, Bitbucket, JIRA, Jupyter, FastAPI, Flask, AWS, Microsoft Azure, Google Cloud Platform (GCP).

Operating Systems and Shells: Linux, Mac Os, Windows, Bash, Zsh, PowerShell.

Big Data Management: Apache Spark, Apache Kafka, MapReduce, HDFS, NFS.

Databases: MySQL, PostgreSQL, MongoDB, DynamoDB, SQL, Data Lakes, ChromaDB, ETL, ELT.