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 in Computer Engineering, specializing in Artificial Intelligence and Machine Learning. Two years of hands-on experience in a dynamic startup environment, leading end-to-end development of cutting-edge AI models and tools. Skilled in transformer architectures, NLP, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG) systems, and computer vision. Experienced in deploying optimized, resource-efficient models both on edge devices and in the cloud.

Professional Experience

Machine Learning Engineer

February 2022 - Present

Knowlix GmbH - Tutzing, Bavaria, 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: Led the supervised fine-tuning (SFT) of an LLM to accurately extract key information from documents, significantly improving data extraction accuracy. Managed end-to-end data curation and selection, implementing a robust data quality assessment mechanism to ensure high-quality inputs prior to training.

Enhanced Vision-Language Model for Complex Document Understanding: Fine-tuned a vision-language model (VLM) to handle complex document structures by effectively integrating visual and textual features. Achieved advanced comprehension of intricate layouts and diverse formats, enabling accurate analysis of highly structured and unstructured document types.

Edge Deployment of Transformers: Created a small, efficient decoder-only transformer model, optimized for deployment on Apple mobile devices using CoreML. Implemented performance optimizations 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. Optimized the model to handle varied document types and complex layouts, enabling precise content categorization.

Serverless REST Endpoint for LLM Interaction: Implemented a serverless RESTful API on AWS Lambda to enable scalable and real-time interaction with a large language model (LLM). Leveraged cloud infrastructure to optimize performance and reduce costs, ensuring secure, efficient model accessibility for end-users.

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, Typescript, 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, RLHF, 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.