Almog Ben Simon

Machine Learning Engineer (MLE) | B.Sc. Data Science & Engineering, Technion. 2 years of production experience at Playtika AI Labs, specializing in scalable GenAI serving infrastructure. Expert in building high-throughput inference systems for self-hosted diffusion models and LLMs, with a strong focus on GPU optimization, Kubernetes deployments, serving Playtika studios, and ML systems engineering. Background in research at Technion.

Experience

2023 -Pres: **Machine Learning Engineer** , **Playtika AI Labs**

- Developed Playtika's enterprise Generative AI Art platform serving pipelines and workflows at scale across studios.
- Led high-throughput inference services, optimized CUDA runtime, improved latency reduction and GPU core utilization.
- Managed LoRA-based plugin system enabling dynamic in-memory model switching.
- Scaled up serving users by improving load balancer bottle-neck and parallel inference using MIG/MPS, optimizing GPU utilization across environments.
- Built automated finetuning infrastructure: end-to-end pipelines for model training, evaluation, and deployment.
- Optimized deployment architecture with k8s, Docker images, and model registry integration across GPU clusters.
- Resolved memory bottlenecks through model compression, quantization, and memory-efficient inference techniques.

2023: **Assistant Researcher , Technion**

- Various Domain shift factors in GNNs and SSL.
- Independently executing extensive research.
- Experienced in custom dataset creation and fine-tuning using libraries as pytorch, graphgym and torch-geometric.

Education

2020-2024: **B.Sc. Data Science & Engineering, Technion**

- Cumulative GPA – **88**

Focusing On: ML, Deep Learning, NLP, Computer Vision, Optimization, Statistic, Distributed Databases, Data Structures & Algorithms, Distributed Programming, Software Engineering.

Biz-Tech (Technion)

- Exclusive entrepreneurship program to develop an early-stage startup concept.
- Admission by winning a Sustainability Startups Contest.

First Semester While Military Service (OpenU)

During a full-time role at the intelligence unit.

2019: **'Geographic Data Analysis' course of Unit 9900**

Focusing on geographic and intelligence data analysis, using tools of Python, SQL, Excel and Intelligence GIS.

2014-2017 : **Full Technological Diploma(CS) at 'Mekif Tet' High School**

almogbensim@gmail.com | (055) 2283025 | Tel-Aviv

LinkedIn https://www.linkedin.com/in/almogbensimon

GitHub https://github.com/almog2065

Projects

- Build an AI-agent model as owner of a taxi business, Deliver the passengers to their destinations in the shortest time.(Python)
- Fine-tuning a LLM (T5) for translating German to English, with hints only in the testset. Combining a dependency parsing model.
- Implementing VAE(discrete,continuous and combined) and Cycle Gan on colored-MNIST dataset (Pytorch).

Military Service

- **Combat soldier - Combat Engineering Corps (1.2 years)**
- **Data Analyst - 9990 unit (1.6 years)**
  - Team leader of operational projects with high responsibility.
  - Working with high-ranking officers and tight deadlines.
  - Completed Unit 9900's Geographic Data Analysis course.
- **Honored by Central Command Intelligence officer (Aluf Mishne)**

Skills

Languages: Python (expert), SQL, Java, Bash, Spark

ML Frameworks: PyTorch, Diffusers, Scikit-learn, OpenCV, NumPy, Pandas

Model Serving & Observability: BentoML, FastAPI, OpenTelemetry, Grafana, Kibana

Infrastructure & DevOps: Docker, Kubernetes (K8s), helm-chart, CI/CD Pipelines

Specializations: Inference Services, Computer Vision, Data Engineering, Deep Learning

Extra

Professional Football player 2006-2016

Intensive training routine with high discipline. Won the state cup and participated in a world tournament in Barcelona (MIC).

Languages

English – Fluent
Hebrew – Native