Vladimir Markov

Senior ML / MLOps Engineer

Skills

Inference optimization

TensorRT, TensorRT-LLM, CUDA, Triton language, SGLang, ONNX, OpenVINO

MLOps & DevOps

Triton Inference Server, Docker, K8s, Helm, CI/CD, Linux, Ansible, Cilium

AI & Machine Learning

PyTorch, TensorFlow, OpenCV, LLMs, CNNs, U-Net

Backend

Python, FastAPI, Pytest, Asyncio, SQLAlchemy, Mypy

C++, CMake, pybind11, OpenGL

Education

2023 - 2025

M.Sc. in Security and Network Engineering

Innopolis University, Russia

01/2022 - 05/2022

Exchange program

Technological University Dublin, Ireland

2019 - 2023

B.Sc. in Computer Science

Innopolis University, Russia

Languages

English - B2

Russian - Native

Contacts

@Markovvn1

linkedin.com/in/markovvn1

markovvn1@gmail.com

markovvn1.me

Summary

Senior ML / MLOps Engineer with 7+ years of commercial software engineering experience, including 4+ years focused on building, deploying, and optimizing production ML systems. Currently working in large-scale fintech, focusing on LLM and speech infrastructure, GPU inference optimization, and cost-efficient model serving. Delivered over $500K in annual infrastructure savings and improved inference performance by up to 4× across high-load STT and LLM systems. Previously built and deployed computer vision solutions for industrial quality control.

Work experience

Senior ML / MLOps Engineer — 07/2023 - present

T-Bank (ex. Tinkoff)

- Improved LLMs serving efficiency by 2× through KV-cache optimization, KV sharing, TensorRT-LLM migration, batching, and request routing improvements.
- Accelerated custom STT models by up to 4× using TensorRT and custom INT8 CUDA kernels.
- Reduced annual infrastructure costs by ~$500K for speech recognition workloads.
- Contributed to optimization and productionization of TTS, STT, and Voice Conversion models.
- Participated in the release of the open-source T-one model: huggingface.co/t-tech/T-one.

Senior Software Engineer — 06/2022 - 07/2023

Mirai Vision

- Early engineer at a computer vision startup, leading development of core systems and helping scale the product to production readiness.
- Led a small engineering team and delivered 5 production-grade ML-based inspection systems in Python for defect detection (PCBs, metal parts, truck frames, rebar, bottles).
- Focused on deployment and optimization of ML pipelines, including running computer vision models on resource-constrained devices (e.g., Raspberry Pi).

Software Engineer — 04/2021 - 06/2022

Center for Oil and Gas Technologies

- Developed a modular web-based VR training simulator.

ML Engineer — 03/2019 - 04/2021

OFTE

- Developed an ML solution for vehicle classification for usedcarsni.com.

Selected projects

ML Engineer / Team Lead — 08/2020 - 04/2021

Eurobot 2021

- Led a 5-member team in preparing for the Eurobot 2021 competition.
- Trained and optimized CNN models for efficient inference on Jetson Nano.
- Gitlab: gitlab.com/inno20_eurobot_full/inno20_eurobot_control.

Achievements

- Speaker at Tinkoff.AI Speech Meetup
- 4 software patents in ML/software systems
- Bronze Medal at “I am a Professional” Olympiad
- Best Innovation Certificate in RoboCup Pacific-Asia 2018