Dmitry Barsukov

Contact Information

Telegram: @riZZZhik (preferred)
GitHub: github.com/riZZZhik
LinkedIn: linkedin.com/in/riZZZhik
Email: riZZZhik@gmail.com

Residence: Moscow, Russia
Open to remote and hybrid full-time jobs

Professional Summary

Experienced Machine Learning Engineer with 6 years in Data Science focusing on Text-to-Speech, Computer Vision, and Deep Learning optimization.

Programming Skills

Languages: Python (advanced), Go (advanced), Tritonlang (average), C/C++ (average), CUDA C++ (average)
Deep Learning Frameworks: PyTorch, TensorFlow, Keras, scikit-learn
Deep Learning Architectures: Transformer, Diffusion, GAN, YOLO, Speech SOTA
Optimization frameworks: TensorRT, torchcompile, Tritonlang, OpenVINO, ONNX Runtime, LiteRT (f.k.a. TensorFlow Lite)
Technical skills: Triton Server, OpenCV, ffmpeg, Torchaudio, WandB, ClearML, Docker, Kubernetes, Git, CI/CD, Observability, Prometheus, Grafana, Linux

Languages

Russian (native); English (advanced)

Employment and Experience

MTS AI June 2022 - Present
Senior Machine Learning Engineer

Development of a Text-to-Speech, Speech-to-Text, and ASR services that outperform leading competitors in the Russian language.

Responsibilities:

- Model deployment using Triton Server, Docker, Kubernetes, Python, and Golang.
- Model optimization for performance and resource efficiency:
  - Model architecture changes
  - TensorRT, tritonlang, torchcompile and OpenVINO
  - Model warmup, quantization, sparsity and pruning
- Research, develop, train and fine-tune new model architectures.

Achievements:

- **0.12** p95 latency and **90** RPS for diffusion model on a single **2g.20Gb** A100 instance.
- Established and automated a version-controlled model deployment process using CI/CD, WandB / ClearML, and Artifactory.
- The development process was established following best practices (e.g., CI/CD, code review, documentation, unit testing, changelog, semantic versioning, etc.).
- Created an automated quality and performance testing in a production-like Kubernetes environment.
- Implemented comprehensive observability with monitoring, logging, and alerting.

Technologies:
Python, Golang, PyTorch, tritonlang, CUDA C++, torchcompile, TensorRT, WandB / ClearML, Triton Server, Observability, Docker + Kubernetes, Git + CI/CD

SIRIN March 2021 - January 2022
Senior Machine Learning Developer

Designed and implemented machine learning service for the automatic opening of car barriers using computer vision.
Achieved **99%** accuracy in recognizing Russian license plates and **90%** accuracy for all other license plates, maintaining a latency of **0.5** on a **4**-core CPU.

Technologies:
Python, PyTorch, OpenCV, Docker + Kubernetes, OpenVINO + Triton Server, Observability (Grafana, Kibana, Prometheus), Git + CI/CD

ITMO University January 2020 - December 2020
Machine Learning Developer

Designed and implemented service for building facade segmentation, managing everything from data collection and preprocessing to model training, evaluation, and deployment.

Technologies:
Python, TensorFlow + Keras, OpenCV, Docker, Git

SPIIRAS August 2018 - October 2020
Middle Machine Learning developer

Designed and implemented facial recognition service, managing everything from data collection/generation and preprocessing to model training, evaluation, and deployment.

Technologies:
Python, TensorFlow + Keras, RealSense DepthCamera, OpenCV, Docker, Git

Education

Higher School of Economics Moscow, Russia (Remote)
B.S., Applied Mathematics and Information Science. September 2023 - Present
GPA: 3.7/4.0