Andrei Panferov

ML Engineer

☎ +7 (926) 431-76-71 | ✉ andrei@panferov.org | BlackSamorez | in BlackSamorez

Experience

Yandex Research
Russia

ML Research Resident — *November 2023 - Present*

- Wrote a first-author paper on *LLM Compression* (see *Publications*)
- Achieved *state-of-the-art* results on *LLM* compression, reducing model size by *87%* with acceptable loss in performance
- Wrote efficient inference kernels using *Triton* and *C++*, speeding up *LLM* inference by up to *320%*
- Integrated the framework into the *transformers* library, enabling low RAM dispatch and reducing instance RAM requirements by *70%*

Eqvilent (High Frequency Trading Fund)
Remote

Software Engineer — *July 2022 - March 2023*

Yandex
Russia

ML Engineer Intern (NLP) — *March 2022 - July 2022*

- Enabled abstract *tabular data* insertion for efficient *map-reduce LLM* inference, speeding up the tabular data processing by 120%
- Increased test coverage of the *map-reduce* inference interface from *0 to 85%* through rigorous unit testing

Terra Quantum AG
Russia

Researcher — *July 2020 - July 2022*

- Researched *quantum algorithms* for business applications
- Optimized *LLM* deployment for chat assistant applications, reducing latency by *40%*

Publications

Extreme Compression of Large Language Models via Additive Quantization
Preprint

Vage Egiazarian*, *Andrei Panferov**, Denis Kuznedelev, Elias Frantar, Artem Babenko, Dan Alistarh — *arxiv.org/abs/2401.06118*

Awards

International Physics Olympiad
Israel

Gold Medal — *Summer 2019*

Education

Moscow Institute of Physics and Technology (MIPT)
Moscow, Russia

Bachelor of Science in Applied Mathematics and Physics — *2020 - 2024*

- Achieved a perfect *5.0/5.0* GPA

Yandex School of Data Analysis (YSDA)
Moscow, Russia

Post-Bachelor's Program in Machine Learning — *2021 - 2023*

- Completed 12 *MSc* level courses. Specialized in *Deep Learning* and *Natural Language Processing*
- Served as a *TA* for the *NLP* course. Prepared a seminar on *Model Compression*, challenged the students to implement *GPTQ*

Open-Source Contributions

tensor_parallel

github.com/BlackSamorez/tensor_parallel

- Developed an open-source *python library* for tensor parallel *PyTorch* models training and inference tightly integrated with *Hugging Face*
- Received more than *400* stars on *GitHub*

🤗 LLaMA implementation for transformers

huggingface.co/docs/transformers/main/model_doc/llama#overview

- Took part in adapting the *LLaMA* model for the *Hugging Face transformers* library, fixing the positional embedding errors

🤗 HuYaLM-100B

huggingface.co/BlackSamorez/HuYaLM-100B-fp16

- Adapted *YaLM-100B* LLM specifically for *Hugging Face transformers*, rewriting the officially published *Megatron-LM* implementation