Andrei Panferov

ML Researcher

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

Summary

I am set to graduate this summer with a Bachelor of Applied Mathematics and Physics from the Moscow Institute of Physics and Technology, where I am currently ranked in the top 15 out of more than 200 students in my cohort. I studied Machine Learning at the Yandex School of Data Analysis, and I have a keen interest in Natural Language Processing, efficient Deep Learning and Federated Learning. My achievements range from securing a gold medal at the International Physics Olympiad to making significant contributions to the open-source community, as well as acquiring work experience in both industry and academia.

Experience

Yandex Research — *Russia*

ML Research Resident — *November 2023 - Present*

- Co-authored a first-author paper on *LLM Compression*. Refer to the *Publications* section for more details

KAUST, Optimization and Machine Learning Lab — *Saudi Arabia*

Research Intern — *July 2023 - September 2023*

- Conducted research, under the supervision of *Prof. Peter Richtárik*, on *distributed optimization*, focusing on *communication compression*
- Authored a first-author paper on *Correlated Quantization*. Refer to the *Publications* section for more details

Eqvilent (High Frequency Trading Fund) — *Remote*

Software Engineer — *July 2022 - March 2023*

Yandex — *Russia*

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

- Refactored and optimized an *LLM inference framework* enabling abstract *tabular data* insertion for efficient *map-reduce* inference
- Increased test coverage of the *map-reduce* inference interface from *0 to 85%* through rigorous unit testing
- Took part in developing a *universal LLM benchmarking solution* adapting *two datasets* for it

Terra Quantum AG — *Russia*

Researcher — *July 2020 - July 2022*

- Researched *quantum algorithms* for business applications
- Developed an NMR spectra analysis toll, allowing for its use for for quantum computations
- 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*

Correlated Quantization for Faster Nonconvex Distributed Optimization — *Preprint*

Andrei Panferov, Yury Demidovich, Ahmad Rammal, Peter Richtárik — *arxiv.org/abs/2401.05518*

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
- Second minor in *Teaching Methods and Pedagogy*
- Working towards my thesis on distributed training of Large Language Models under the supervision of *Prof. Alexander Gasnikov*

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*
- Contributed significantly to *Open-Source* (see Open-Source Contributions)
- 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

NLP Bot Project

github.com/BlackSamorez/ebanko

- Designed an automatic data collection system to extract thousands of dialogues from internet forums, refined the collected data using a pretrained sentiment analysis *BERT* model and published them it a dataset
- Fine-tuned a *GPT-2* model for *chatbot* purposes on the refined dataset and deployed it as a Telegram *bot*
- Published an article on *Habr* (IT social network) about the project, reaching the daily top-1 in the *ML* section