Andrei Panferov

ML Researcher

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

Experience

KAUST, Optimization and Machine Learning Lab *Saudi Arabia*
RESEARCH INTERN *July 2023 - September 2023*

- Conducted research under the supervision of *Prof. Peter Richtárik*
- Derived theory and ran experiments on *distributed optimization*, focusing on *communication compression*
- Submitted a first-author paper to an upcoming conference, with the preprint soon to be on arxiv (see Publications)

Eqvilent (High Frequency Trading Fund) *Remote*
SOFTWARE ENGINEER *July 2022 - March 2023*

Yandex *Moscow, 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 from *0 to 85%* through rigorous unit testing
- Took part in developing a *universal LLM benchmarking solution* adapting *2 datasets* for it

Terra Quantum AG *Moscow, Russia*
RESEARCHER *July 2020 - July 2022*

- Researched *quantum algorithms* for business applications
- Developed an NMR spectra analysis toll, allowing for it's use for for quantum computations
- Optimized *LLM* dedployment for chat assistant applications, reducing latency by *40%*

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*

Yandex School of Data Analysis (YSDA) *Moscow, Russia*
INDUSTRY-ORIENTED PROGRAM IN MACHINE LEARNING *2021 - 2023*

- Completed 12 *MSc* level courses. Specialized in *Deep Learning* and *Natural Language Processing*
- Rigorously contributed to *open-source* (see Open-Source Contributions)
- Served as a *TA* for the *NLP* course. Designed a homework on *Model Compression* and graded a class of 200+ students on it

Publications

Correlated Quantization for Faster Nonconvex Distributed Optimization *KAUST, Saudi Arabia*
ANDREI PANFEROV, YURY DEMIDOVICH, AHMAD RAMMAL, PETER RICHTÁRIK *Under Review*

Research Interests

- Efficient Deep Learning

- Natural Language Processing

- Federated Learning

- ML Systems

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 inregrated with *Hugging Face*
- Recieved more than *400* stars on *GitHub*

LLaMA implementation for transformers

HUGGINGFACE.CO/DOCS/TRANSFORMERS/MAIN/MODEL_DOC/LLAMA

- Took part in adapting the *LLaMA* model for the *Hugging Face transformers* library, fixing the positional embedding errors and optimizing past key-value handling

HuYaLM-100B

HUGGINGFACE.CO/BLACKSAMOREZ/HUYALM-100B-FP16

- Adapted *YaLM-100B* LLM specifically for *Hugging Face transformers*, rewriting the officialy 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