Vladimir Baikalov

R&D Research Engineer / PhD Student

✉: nonameuntitled159@gmail.com
in: linkedin.com/in/noname-untitled
: github.com/NonameUntitled
: t.me/noname_untitled

Work experience

Team lead R&D RecSys Engineer/Yandex Technologies — June 2024 - Present

Relevant areas: RecSys, Information Retrieval, Highload, Deep Learning

Leading a team of three engineers. Researching the application of e2e models instead of multi-stage classic recommender system pipeline. Investigating limitations and possibilities of applying semantic IDs for generating retrieval for stream music domain.

Deployed retrieval and ranking models for Yandex.Lavka (e-commerce service). This model increased **GMV by 2.1%**.

Publications:

Short paper «**Correcting the LogQ Correction: Revisiting Sampled Softmax for Large-Scale Retrieval**» – Improving the LogQ correction presented at **Google, Deepmind**. Paper was accepted at RecSys'25

Full paper «**Scaling Recommender Transformers to One Billion Parameters**» – demonstrating scaling laws in recommender systems and presenting novel pretraining idea.

Reproducibility paper «**Yambda-5B — A Large-Scale Multi-Modal Dataset for Ranking and Retrieval**». The largest dataset of streaming music data. Paper was accepted at RecSys'25.

Competitions:

With the team took the 4th place at RecSys'25 challenge. The solution «**Blending Sequential Embeddings, Graphs, and Engineered Features**». Paper was accepted at RecSys'25.

Machine Learning Engineer/Google, YouTube — January 2023 - June 2024

Relevant areas: Transformers, Highload, Large Language Models (LLMs)

Developed video games detection algorithm for shorts. The final model achieved **4% recall boost without precision drop** and allowed to **conduct fine-tuning more frequently**. **Designed an end-to-end system** for serving this classifier, the solution is now being hosted by YouTube Gaming team.

Worked on deployment of LLMs in production including improvement of **fine-tuning and distillation** pipelines.

Deep Learning Engineer/Yandex Technologies — August 2021 - December 2022

Relevant areas: RecSys, Information Retrieval, Highload, Deep Learning

Transferred an ML model from experimental setup (Python, PyTorch) to production framework (C++, YNMT). Supported weekly continuous fine-tuning process for **Yandex.Ads. It is still being applied in production.** This model increased **GMV by up to 1.5%** and **CTR by up to 5%**. The result is verified by AB tests and experiments.

Applied encoder-based models for improving personalized ads and search recommendations. The current solution **boosts production metrics by up to 6%**.

Implemented multiprocessing Python package (YtReader) for fast and convenient data preprocessing. The final solution reduced the time required for model training/evaluation **by up to 5 times**.

Machine Learning Engineer/Huawei R&D Dept. — March 2020 - July 2021

Relevant areas: Object Detection/Tracking, Digital Sound Processing, Optical Character Recognition.

Algorithm for vehicles trajectories prediction using radar data only. This approach now is being used in a real-world application. Proposed solution **outperforms the previous algorithm by 44%**.

Algorithm for the knuckle-knock sound pattern detection. Proposed architecture achieved **90%** Precision and **95%** Recall. With the usage of touch sensors precision was improved up to **94%**.

Education

PhD student/ITMO University, RecSys dept. — September 2024 - Present

Working on improving retrieval models by applying enhanced user/item representations.

Master degree/Skoltech, DS major — September 2021 - July 2024

GPA: 4.5 / 5.0. Thesis project: End-to-end Graph-Sequential Representation Learning for Accurate Recommendations.

Publications: Short paper was accepted at **TheWebConf'24 conference (A*-tier)**, Singapore.

Received certificate of achievement **"Best Research Thesis"**.

Bachelor degree/ITMO University, CS major — September 2017 - August 2021

GPA: 4.7 / 5.0. Developed multi-agent policy-based algorithm **REM (Reinforce, Embedding, Monte-Carlo)** for baggage handling system.

Publications: «**Multi-Agent Deep Reinforcement Learning-Based Algorithm For Fast Generalization On Routing Problems**» was published at the **YSC 2021 conference**. This project was done in collaboration with **Aalto University, Finland**.