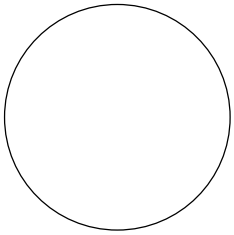
AIRI

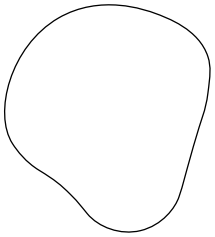
2023

Annual report



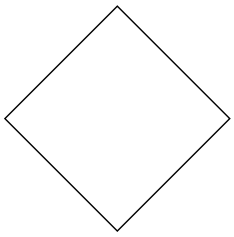
Nature

Constant movement and evolution, regularity and systematicity, laws and rhythms, chaos and life



Human

A living organism, part of nature, capable of thought, feeling and speech, imperfect



Technology

Human creation and its extension, logical and predictable, utilitarian, designed to help

AIRI

The unity of nature, human beings and technology

Table of Contents

CEO’s statement	4
AIRI Mission	5
AIRI Values	6
Focus areas	7
Key results	8
Research group leaders	10
Management	12
Scientific results	13
Publications	40
Events and special projects	62
Partnerships and collaborations	74
Contact information	76

The Institute’s third year

Ivan Oseledets

CEO

In 2023, the Institute solidified its position as one of Russia’s leaders in artificial intelligence and is on track to become the top publisher at A/A* conferences in the future.

Our main objective for the upcoming years is to develop multimodal, multi-agent systems of AGI. This year, our teams made significant contributions to several large models, including Kandinsky, Russia’s first text-to-video model KandinskyVideo, and Russia’s first multimodal model, OmniFusion.

Many teams in the country are currently working on artificial intelligence, such as Skoltech, Higher School of Economics, MIPT, ITMO, Innopolis, and ISP RAS. However, AIRI is unique in its primary goal of creating and developing strong artificial intelligence. I believe we can achieve it, as technology has reached the necessary level, and we have the strongest AI teams in the country.

In 2023, we welcomed new colleagues. Alexander Panchenko is now the head of the NLP group, and Dmitry Dylov is leading the ‘Medical AGI’ laboratory. We also have plans to create additional teams in the future.

The Institute’s core values remain centered on human-centricity, responsibility, and contribution to the future. We continue to actively pursue educational activities, bringing together a community of enthusiasts to discuss pressing issues in the subject area.

In 2023, our portfolio has produced significant scientific findings on both core and applied aspects of AI. Detailed information about them can be found in this report.

AIRI Mission

To create universal AI systems that solve real-world problems

The main goal of AIRI is to find opportunities to use artificial intelligence for solving complex scientific, social and economic problems. The Institute's researchers conduct both core and applied research with the aim of making significant advancements in the field of AI and its applications. Their work contributes to shaping the global research agenda.

AIRI Values

Human-centricity

Scientific freedom

Responsibility
& contribution

Openness
& transparency

Collaboration
& partnership

Focus areas

Research

Conduction of breakthrough research in the field of artificial intelligence and work towards the formation of a global center of expertise

Contribution to the development of artificial intelligence

Participation in the global development of artificial intelligence through the creation, development and support of open source projects

Scientific and industrial partnerships

Development of partnerships with scientific organizations, industry and government, development and commercialization of technologies in the field of artificial intelligence

Laboratories

Cooperation with institutes, universities and industrial partners to launch joint research laboratories in the field of artificial intelligence

Popularization of AI

Holding specialized conferences and events, creating and supporting competitions, promoting AI technologies

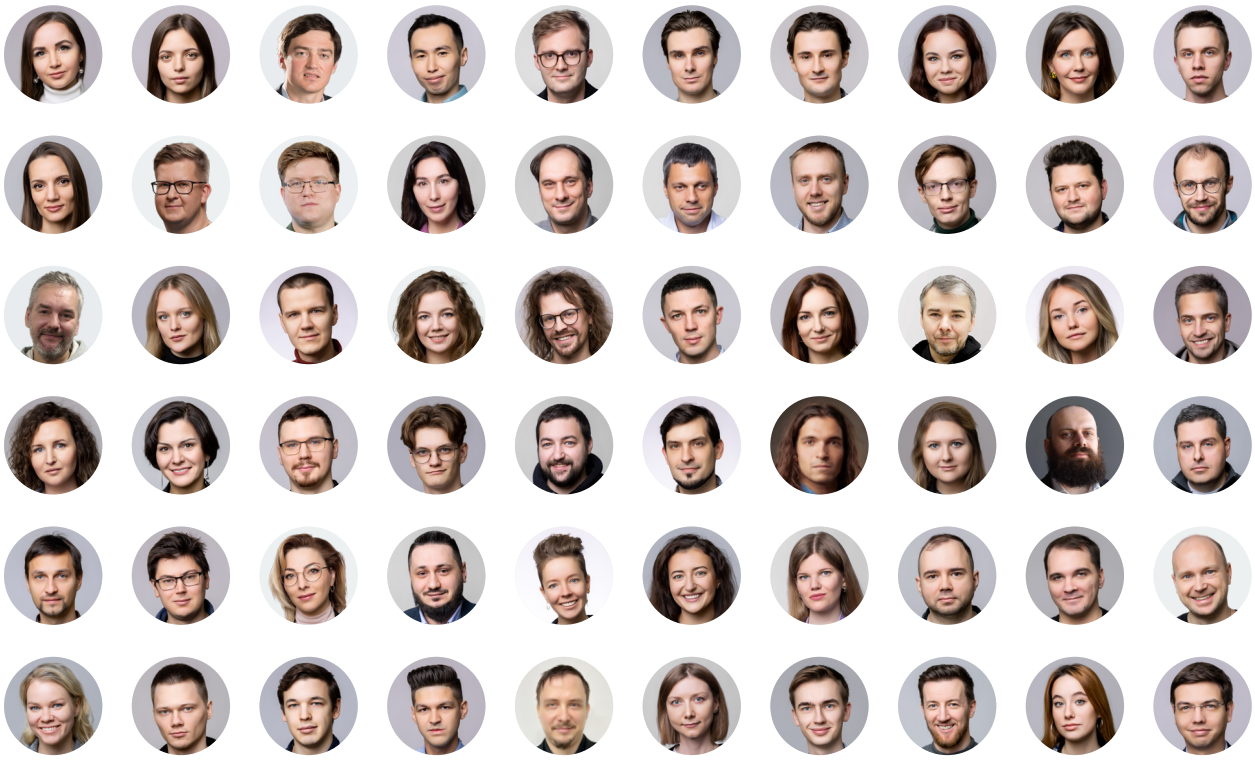
Key results

146
publications

30
conference papers (A*)

43
journal papers (Q1)

8
conference papers (A)



Research group leaders

Artem Shelmanov

Weakly-Supervised NLP

Artur Kadurin

DL in Life Sciences

Evgeny Frolov

Personalization Technologies

Oleg Rogov

Reliable and Secure Intelligent Systems (RSI)

Aleksandr Panov

Neural-Symbolic Integration

Dmitry Vetrov

Probabilistic Learning

Olga Kardymon

Bioinformatics

Andrey Kuznetsov

FusionBrain

Research group leaders

Alexander Panchenko

Computational Semantics

Elena Tutubalina

Domain-specific NLP

Alexey Ossadtchi

Neurointerfaces

Semen Budennyy

New Materials Design

Ilya Makarov

Industrial AI

Evgeny Burnaev

Learnable Intelligence

Dmitry Dylov

“AGI Med” Lab

Management

Ivan
Oseledets

CEO

Manvel
Avetisyan

Director of applied
projects development

Maksim
Kuznetsov

Head of Project
Management Office

Anton
Rizaev

CFO

Alexandra
Broytman

Marketing
and Communications Director

Maria
Marakhovskaia

HR Director

Yuliya
Nikitina

General Counsel

Stepan
Mamontov

Chief developer, Head
of Research Support department

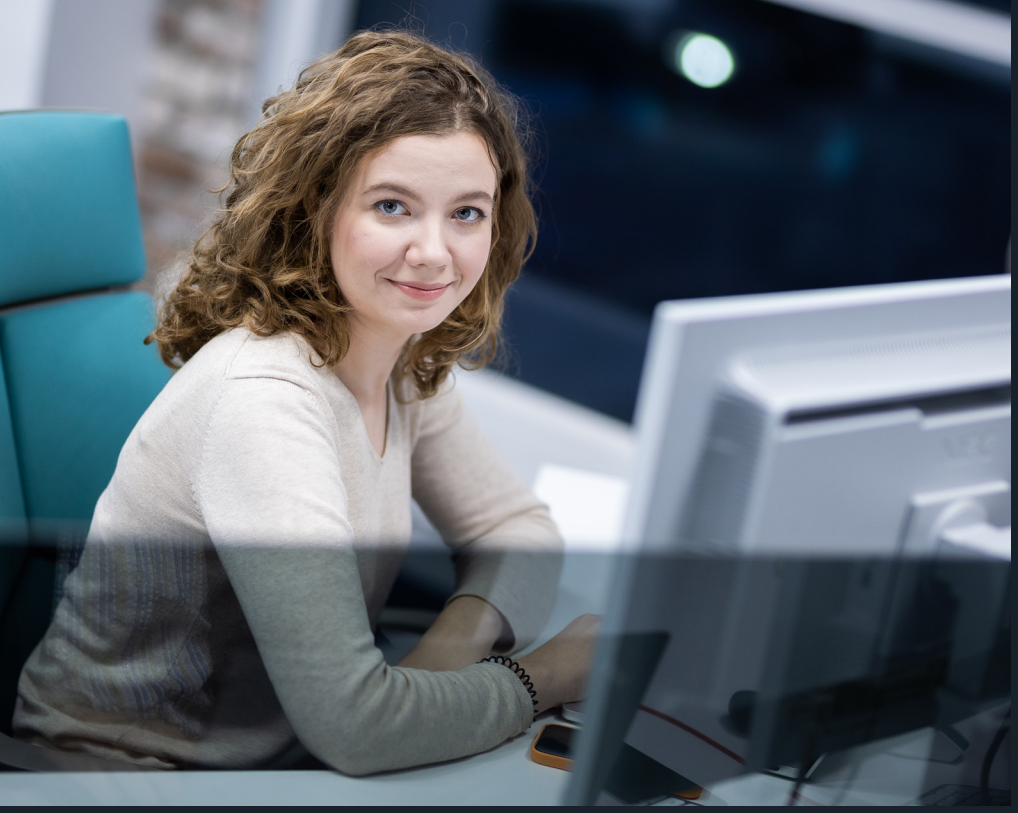
Konstantin
Katanov

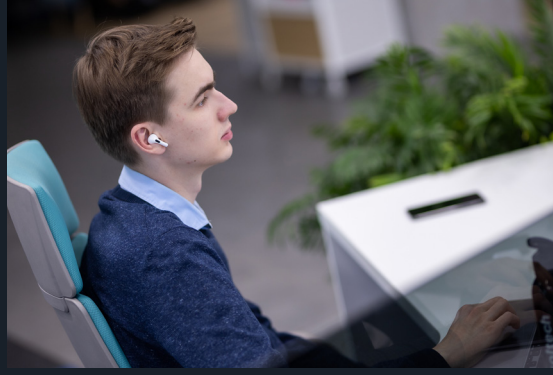
Head of IT department

Olga
Surovegina

Science & Technology
Partnerships Director

Scientific results





Main results

Andrey Kuznetsov

Head of FusionBrain Research,
Lead Researcher

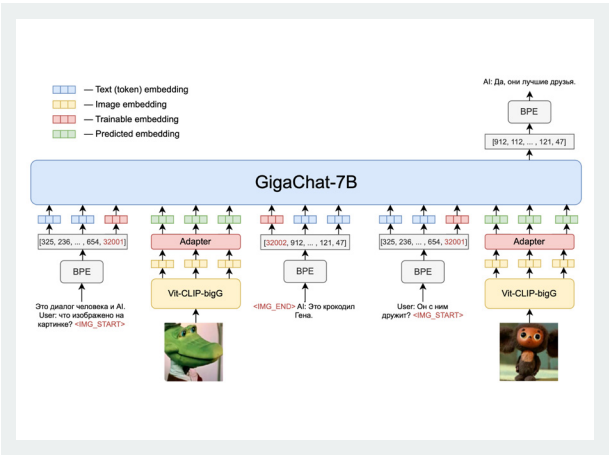
Denis Dimitrov

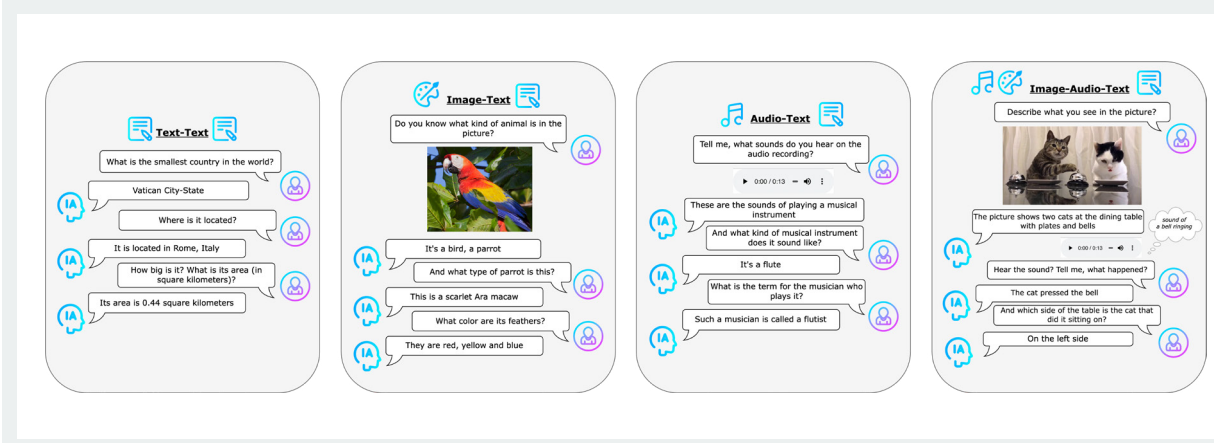
Scientific Consultant
of FusionBrain

OmniFusion

The development entails the creation of the first Russian multimodal dialogue model, capable of engaging in conversations with users based not only on text but also incorporating images.

The model demonstrates proficiency in image comprehension, enabling it to generate conversation based on them, respond to various questions and provide detailed descriptions related to scenes and objects. Leveraging the GigaChat-7B and Mistral-7B language models (for publication in open source), the training process was conducted in two stages: first, involving large-scale multimodal pre-training for feature alignment, followed by fine-tuning on downstream tasks.



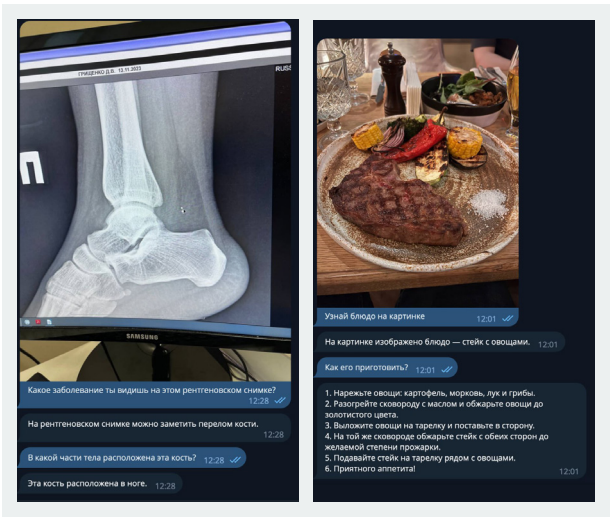


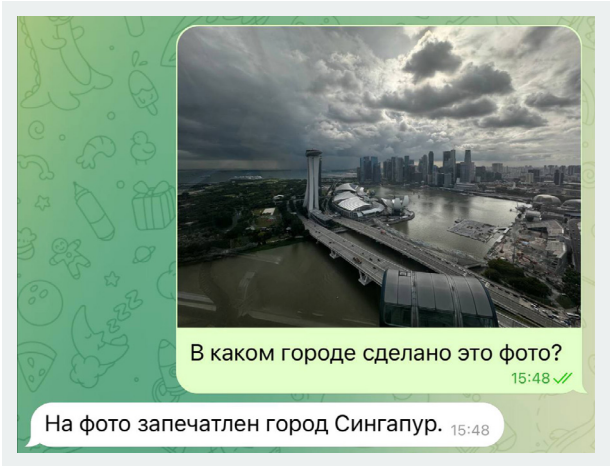
Project Partners

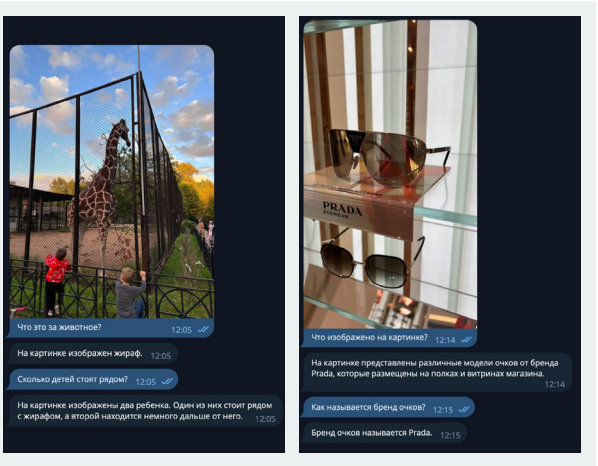
SBER AI SBER DEVICES

Comparison with the open-source LLaVA model across 10 benchmarks revealed that the OmniFusion model matches or surpasses its performance, despite utilizing a more lightweight language model with 7 billion parameters compared to LLaVA's 13 billion. This means that the model is both more cost-effective and faster.

The current version of the model is already working reliably in English and is undergoing training to achieve fluency in Russian, in order to be accessible to users. Additionally, the team is in the process of preparing a scientific publication that details the creation of OmniFusion. Current efforts involve testing the model's integration into GigaChat, while also exploring its potential use as a standalone feature.







Andrey Kuznetsov

Head of FusionBrain Research,
Lead Researcher

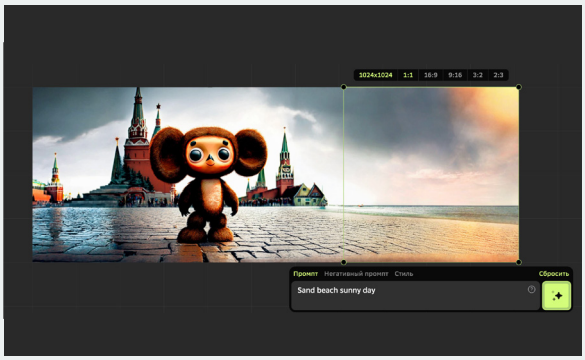
Denis Dimitrov

Scientific Consultant
of FusionBrain

Kandinsky 2.X, 3.0

A research and experimentation plan were devised to create a novel diffusion model for image synthesis from text descriptions in multiple languages.

The research covered various experiments with architecture that focused on the image prior component and assessed its impact on the generation quality.
As a result of the research, an article was prepared and accepted for presentation at the EMNLP conference (Core A*).
The article also ranked first among the Daily Papers on the Hugging Face, surpassing articles from Google DeepMind and Carnegie Mellon University.



Kandinsky Video

The research included the creation of the first in Russia end-to-end model for video synthesis from text descriptions based on a diffusion model for frame generation.

This allows for controlling video plot creation, ensuring smooth object motion by filling frames up to 30 FPS, and enhancing the visual quality of the synthesis. The article describing the model is currently undergoing peer review. It also ranked second among the Daily Papers on the Hugging Face, second only to Yann LeCun's article on a new multimodal benchmark.

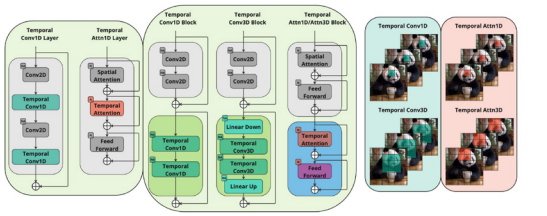
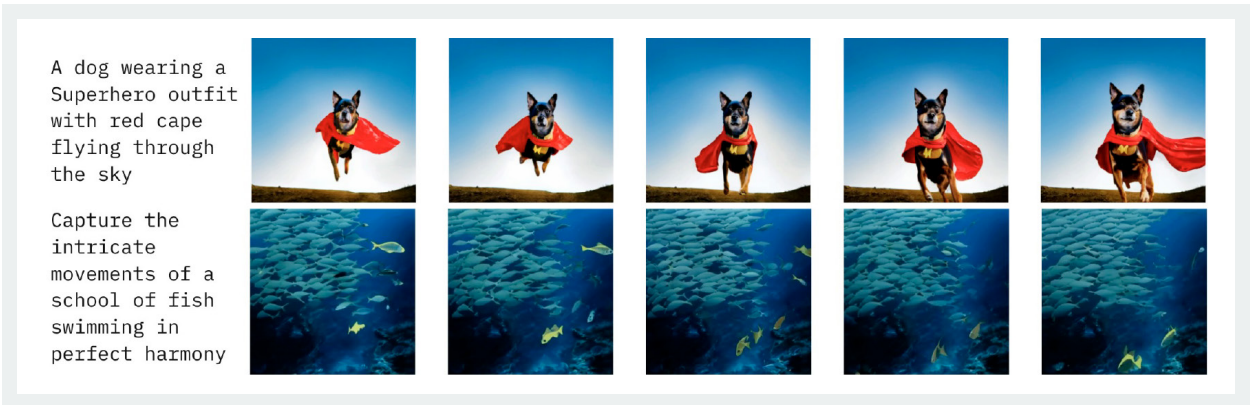


Figure 3. **Comparative schemes for temporal conditioning.** We examined two approaches of temporal components use in pretrained architecture of T2I U-Net from Kandinsky 3.0 [1] – the traditional approach of mixing spatial and temporal layers in one block **(left)** and our approach of allocating a separate temporal block **(middle)**. All layers indicated in gray are not trained in T2V architectures and are initialized with the weights of the T2I Kandinsky 3.0 model. NA and N in the left corner of all layers correspond to the presence of prenormalization layers with and without activation, respectively. For different types of blocks we implemenied different types of temporal attention and temporal convolution layers, **(left)**. We also implement different types of temporal conditioning. One of them is simple conditioning when pixel see only itself value in different moments of type (1D layers). In 3D layers pixels can see the values of its neighbors also in different moments of time **(right)**.



Major results

Advances in natural language processing (NLP) technologies

Groups led by Alexander Panchenko, Artem Shelmanov, and Elena Tutubalina obtained several new results

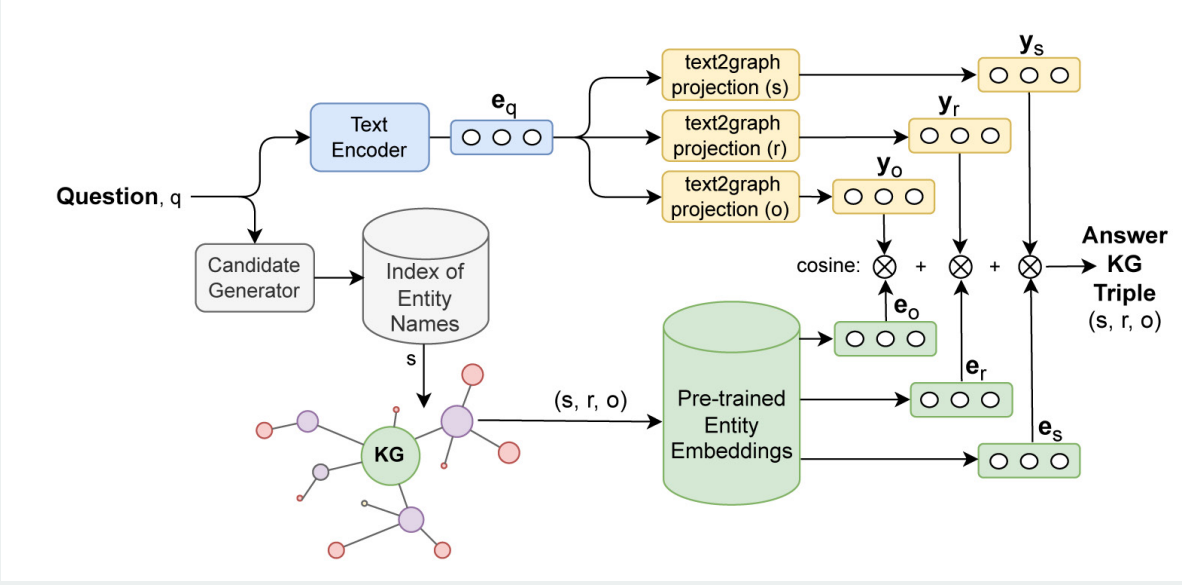
Alexander Panchenko
Leading Researcher

Artem Shelmanov
External Scientific Advisor

Elena Tutubalina
Scientific Consultant

New methods have been developed for question-answering systems using knowledge graphs, which allow for better answers to simple questions than ChatGPT.

The developed methods allow for answering questions using knowledge graphs and large language models. On multiple datasets, the methods show improvements over strong baseline approaches, including ChatGPT.



New computationally efficient methods have been presented for evaluating uncertainty in language model predictions.

The developed approach combines two state-of-the-art methods to achieve significant quality gains in selective classification tasks. Such a hybrid approach has shown its effectiveness for tasks with high levels of ambiguity, such as determining the degree of toxicity or sentiment in texts. Additionally, despite not requiring large computational resources, this method is capable of outperforming methods such as deep ensembles, which require multiple increases in computational resources.

The LM-Polygraph library has been developed for evaluating uncertainty in LLM, including models deployed as services.

Uncertainty estimates allow for detecting low-quality model responses (e.g. hallucinations) and refusing to display them to users.

A new benchmark has been developed based on the PAUQ dataset.

Four text-to-SQL partitions of our PAUQ dataset, called Shifted PAUQ, were proposed to evaluate compositional and multilingual generalization of SoTA text-to-SQL models. Metrics for evaluating model compositionality were developed using context-free grammars.

Research was conducted on uncertainty estimation methods in conjunction with debiasing methods.

The most promising debiasing approaches were identified, which have a minimal negative effect on uncertainty estimates.

A new framework, Vote'n'Rank, was proposed, which includes eight methods for aggregating model metrics based on social choice theory.

In addition, practical recommendations for using the framework have been presented based on the theoretical properties of results aggregation methods and potential application scenarios.

Rank	σ^{am}	σ^{gm}	σ^{og}	Copeland	Minimax	Plurality	Dowdall	Borda
1	91.18	90.89 ↕0	0.074 ↕0	29.00 ↕0	0 ↕0	2.00 ↕0	4.95 ↕0	260.50 ↕0
2	91.07	90.78 ↕0	0.075 ↑4	25.00 ↑1	−5.50 ↑1	2.00 ↑13	4.08 ↑13	256.00 ↕0
3	90.88	90.56 ↕0	0.076 ↓1	24.00 ↓1	−6.00 ↑1	1.50 ↕0	3.82 ↕0	247.50 ↕0
4	90.86	90.48 ↕0	0.076 ↕0	22.00 ↑3	−6.50 ↓2	1.00 ↑1	3.41 ↕0	241.50 ↕0
5	90.74	90.44 ↕0	0.077 ↕0	22.00 ↑10	−7.00 ↑2	1.00 ↓3	3.27 ↓3	233.50 ↑1
6	90.66	90.34 ↕0	0.078 ↑1	22.00 ↓2	−7.00 ↑9	0.50 ↕0	2.57 ↓1	229.50 ↑1
7	90.48	90.11 ↕0	0.082 ↑3	16.00 ↓1	−7.00 ↓1	0.00 ↓3	2.55 ↕0	220.50 ↓2

Table 3: Results of re-ranking the GLUE benchmark. Changes in the system ranks are depicted with arrows, whilst the superscripts denote scores assigned by the aggregation procedure. Notations: =HUMAN; =ERNIE; =STRUCTBERT+CLEVER; =DEBERTA+CLEVER; =DEBERTA/TURINGNLRV4; =MACALBERT+DKM; =T5; =ALBERT+DAAF+NAS; =FUNNEL. The superscript values stand for the voting rules' scores, whilst the subscript values indicate changes in the ranking positions. ↑ x means up x positions, ↓ x means down x positions, ↕ means no changes.

Probabilistic Learning

The group led by Aibek Alanov and Dmitry Vetrov obtained important results

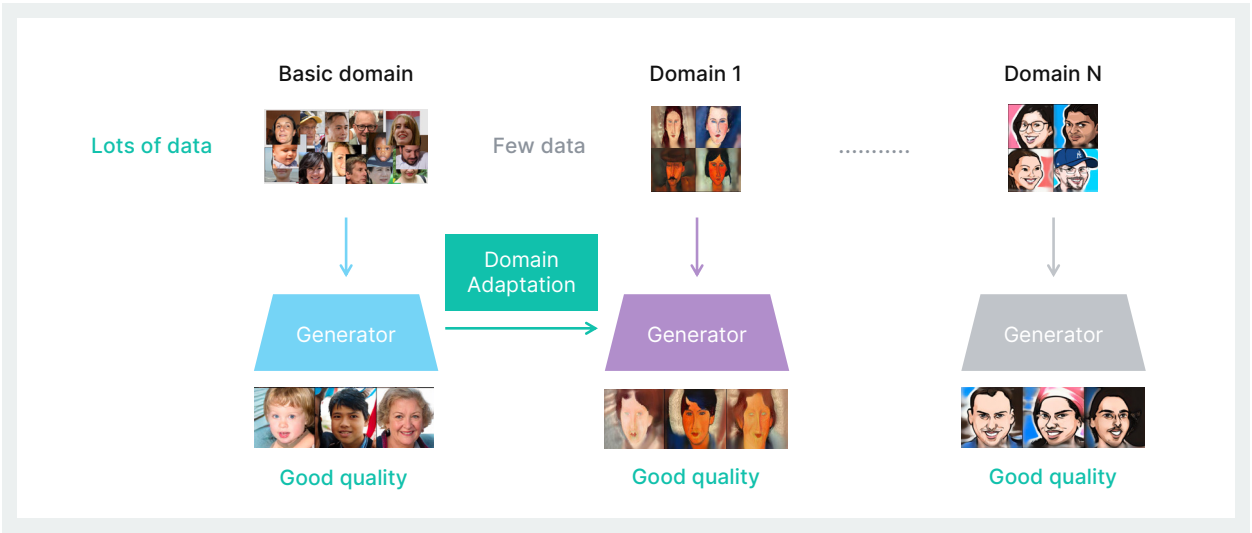
Dmitry Vetrov
Leading Researcher

Aibek Alanov
Researcher

Domain adaptation of GAN for face generation has been made 5000 times faster.

As a result, an article was published at the ICCV conference. This work examined the area of generating synthetic (indistinguishable from real) images from a given domain (photos of faces, houses, cats). More precisely, the problem of domain adaptation is solved for the sota model StyleGAN2.

The essence of the problem is to match each synthesized image with a stylized copy from another domain (drawing, watercolor, cartoon style). All previous (including sota) methods used additional training of a full StyleGAN2 copy — our proposed method solves the problem using 5000 times less memory during training and inference, and at the same time achieves comparable quality to existing approaches.



Improving diffusion models.

A new model has been presented that can be applied to various distributions in the exponential family, making it useful for modeling data on manifolds with constraints, such as the unit sphere or positive semi-definite matrices (NeurIPS 2023).

Improving generative models for sound generation.

Two papers were co-written with colleagues from the Samsung AI Center and accepted at the ICASSP 2023 and InterSpeech 2023 conferences.

A method has been developed for manipulating images using the generative model StyleGAN, and a paper has been submitted to the CVPR 2024 conference.

In this work, the problem of inversion of real images for the StyleGAN model is solved. The proposed solution completely outperforms all existing (including SOTA) methods according to objective reconstruction metrics. The distinctive feature of the solution is its high ability to edit real images and its ability to be applied to extreme out-of-domain examples.

A method has been developed for editing real images using text-to-image diffusion models, and a paper has been submitted to the CVPR 2024 conference.

The proposed method outperforms all major (including SOTA) methods for editing based on real user feedback and on average by metrics. The distinctive feature is its robustness to hyperparameter selection — that is, the method performs well when selected from a sufficiently large range.

A method has been developed for efficiently solving the HairSwap problem using the generative model StyleGAN, and a paper has been submitted to the CVPR 2024 conference.

This work considered the challenging task of transferring hairstyles based on an input photo. The main difficulty of this task is achieving a natural and realistic result when transferring hair. The proposed method solves this problem without optimization, which significantly speeds up its operation while achieving better quality compared to other approaches.

Learnable Intelligence

The group led by Evgeniy Burnaev received new results in optimal transport and its applications

Evgeniy Burnaev
Leading Researcher

New methods for building generative models based on optimal transport and diffusion processes (results published in 3 papers at ICLR, 2023 and 4 papers at NeurIPS, 2023).

Notably, the papers at ICLR, 2023 (Notable-top-25% paper) and NeurIPS, 2023 (Oral talk, 3% of accepted papers). The first work proposes a method for constructing a generative model based on a new efficient algorithm for solving the weak optimal transport problem. The second work investigates the equivalence between the continuous optimal transport problem in the entropy formulation and the problem of finding the diffusion Schrödinger bridge; a scalable neural network algorithm is proposed for solving both problems. The developed method allows for efficient solutions to unpaired learning tasks, such as enhancing image resolution.

New methods for analyzing the structure of multidimensional data based on topological data analysis (paper at ICLR, 2023).

Based on a topological approach to estimating dimensionality, an approach to detecting fake texts generated by large language models, such as GPT4, is proposed (paper at NeurIPS, 2023).

A new method was developed based on the proposed approaches to optimal transport calculation, which can work in conjunction with a set of existing offline reinforcement learning methods.

In particular, in combination with the recently proposed SOTA method — ReBRAC, developed by Tinkoff Bank's team, which provides acceleration and improved accuracy of solutions by approximately 40%, the proposed approach allows for an additional improvement of approximately 20% on tasks in the AntMaze environment.

It has been shown that the set of topological features previously developed by the authors could be applied to a completely new type of data — speech — when solving standard classification tasks and, in some cases, achieves SOTA results (paper at Interspeech, 2023).

New methods for 3D computer vision and 3D shape reconstruction.

A dataset of diverse multimodal 3D data has been constructed, which allows models to be trained for 3D shape reconstruction. This result will provide the scientific community with quality data for training various 3D computer vision models (2 papers at CVPR, 2023).

AIRI scientists have developed a range of methods for optimal transfer of data from one domain to another using neural networks.

Such methods are currently actively being researched in the scientific community and are often used in generative modeling, image processing, and biological data analysis.

The theory of optimal transport, a significant contribution to the development of which was made by the Soviet mathematician Leonid Kantorovich back in the 20th century, underlies these new methods. That is why the proposed methods based on neural networks have a common name “neural optimal transport”. These new approaches have greater interpretability and theoretical justification than many existing alternative approaches to data domain translation, which are usually based on heuristic principles and do not have a rigorous theoretical basis.

Fig. (a)

Fig. (b)

Examples of the developed neural network algorithm’s work on three domain transfer tasks with unpaired training samples. Fig. (a) — stylization of a photorealistic portrait into anime (first two rows) and generation of a building image from a given landscape image (second two rows). Fig. (b) — generation of shoe images based on the style specified by the input bag image (rows 2-4 show diverse generation results).

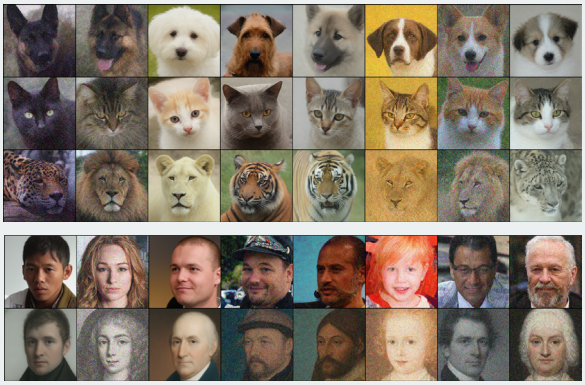
Computational Intelligence

The scientific team led by Ivan Oseledets obtained several results

Ivan Oseledets
CEO

It has been shown that in some cases, diffusion models implement optimal transport between the source distribution and the normal distribution, and a hypothesis has been formulated and tested in the general case.

It has been demonstrated that two diffusion models trained on separate datasets give “close” latent codes for “close” images. The paper has already been cited 18 times in 2023, according to Google Scholar. The paper was published at ICLR 2023.



A series of works have been carried out to develop machine-learning methods for mathematical modeling.

A method for fast dataset augmentation by using equation properties (data augmentation) to train neural operators has been proposed. A paper was published at ICML 2023.

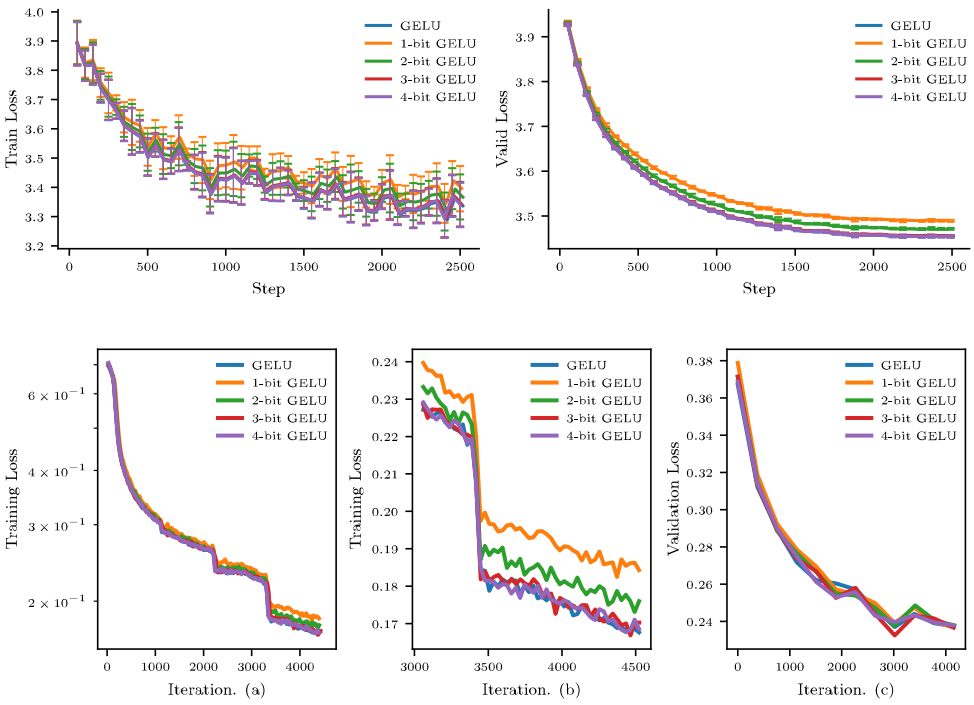
Equation	Model	simple datasets			complex datasets		
		×	√	g	×	√	g
Convection-diffusion	FNO	0.067	0.048	28%	0.510	0.418	18%
	DeepONet	0.675	0.567	16%	—	—	
	DilResNet	0.023	0.010	56%	0.312	0.225	28%
	MLP	0.094	0.050	49%	0.566	0.496	12%
	U-Net	0.069	0.031	55%	0.419	0.364	13%
	SNO	0.086	0.066	23%	0.416	0.373	10%
Elliptic alpha	FNO	0.066	0.036	46%	0.306	0.207	32%
	DeepONet	—	0.826		—	—	
	DilResNet	0.105	0.021	80%	0.160	0.133	17%
	MLP	0.088	0.053	40%	0.322	0.253	21%
	U-Net	0.093	0.070	25%	0.386	0.194	50%
	SNO	0.082	0.050	39%	0.251	0.209	17%
Elliptic beta	FNO	0.034	0.021	38%	0.181	0.126	30%
	DeepONet	—	0.832		—	0.946	
	DilResNet	0.099	0.022	78%	0.089	0.062	30%
	MLP	0.069	0.035	50%	0.238	0.138	42%
	U-Net	0.070	0.067	4%	0.170	0.143	16%
	SNO	0.068	0.038	44%	0.187	0.144	23%
Wave	FNO	0.200	0.159	21%	0.650	0.628	3%
	DeepONet	—	—		—	—	
	DilResNet	0.053	0.048	9%	0.43	0.38	12%
	MLP	0.313	0.295	6%	—	0.99	
	U-Net	—	—		0.57	0.52	9%
	SNO	0.37	0.37	0%	—	—	
		v^1			v^2		
Navier-Stokes	FNO	0.005	0.003	40%	0.022	0.010	55%
	UNet	0.019	0.09	53%	0.069	0.037	46%
	DilResNet	0.021	0.015	29%	0.073	0.045	38%
	MLP	0.082	0.066	38%	0.082	0.066	20%
	SNO	0.004	0.003	25%	0.013	0.008	38%

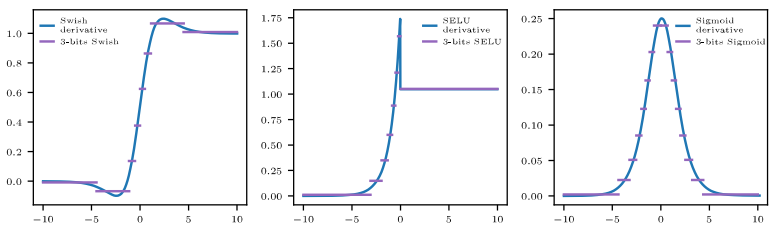
A series of works have been carried out to develop tensor methods: a constructive approach to building tensor approximations for special functions.

In particular, a significant acceleration of cooperative games calculation has been achieved, and an optimal algorithm for computing permanents has been obtained. This algorithm received further development by the end of 2023 to accelerate the modeling of bosonic samplers, which play an important role in photon quantum computers. A new tensor optimization method of the “black box” type — PROTES — has been developed, which significantly outperformed analogs, including the Nevergrad (Meta*) package on several benchmarks. A paper was published at NeurIPS 2023. A new method for optimizing functions represented in low-rank format has been developed. A paper was published in SISC (Q1).

A memory reduction method for training neural network models by quantizing activations in the backward pass (FewBit method) has been developed.

Memory reduction without loss of accuracy was 5–10%. A paper was published at ICML 2023.





* The organization is recognized as extremist and its activities are banned on the territory of the Russian Federation

Neural Symbolic Integration

The group led by Aleksandr Panov obtained the following results

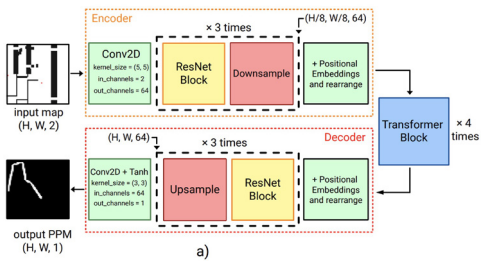
Alexander Panov
Principal Research Scientist

A method for integrating transformer neural network models and heuristic search algorithms for the pathfinding problem on a graph of a regular decomposition (planning a trajectory along a grid graph) is proposed.

The work was published at the AAAI 2023 (Core A*) conference.

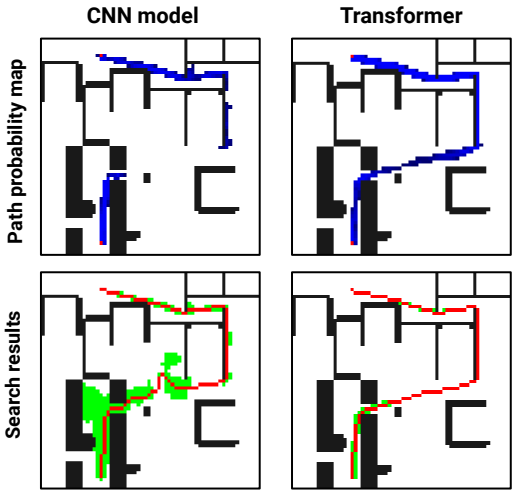
New methods for planning the trajectory of a mobile agent in an environment with moving obstacles are proposed, taking into account the kinodynamic constraints of the control object, including methods that combine learnable policies for obstacle avoidance strategies and classical planning algorithms (heuristic search, planning based on random sampling).

The works were published at the AAAI 2023 conference (Core A*), RA-L journal (Q1, Top-1 in Robotics).



a)

	Optimal Found Ratio (%) ↑	Cost Ratio (%) ↓	Expansions Ratio (%) ↓
A*	100	100	100
WA*	40.66	103.52 ±4.85	44.43 ±25.92
Neural A*	29.82	104.90 ±6.56	52.30 ±30.47
A*+HL	79.11	100.27 ±0.62	80.50 ±74.40
WA*+CF	**85.40**	100.25 ±1.13	36.98 ±21.18
FS+PPM	82.97	**100.24 ±0.74**	26.36 ±21.08
GBFS+PPM	83.02	100.25 ±0.90	**23.60 ±18.34**



Applied results

NLP

Groups led by Alexander Panchenko, Artem Shelmanov, and Elena Tutubalina obtained the following results

Alexander Panchenko
Leading Researcher

Artem Shelmanov
External Scientific Advisor

Elena Tutubalina
Scientific Consultant

- A new method for assessing uncertainty in predictions has been implemented in a medical diagnostic system developed by Sber AI Lab, which is deployed in clinics in Moscow.

- Development of methods for cross-lingual and multilingual text detoxification. As part of this work, methods were presented for “translating” toxic text messages containing profanity into equivalent neutral messages. Our study was the first to test detoxification methods in a multilingual and cross-lingual setting. The latter implies that two tasks are simultaneously solved: translation and detoxification of the text.
The results of the method were presented at the AACL-IJCNLP-2023 conference in Indonesia. Additionally, an application was accepted to hold an international competition on Multilingual Text Detoxification at the CLEF-2024 forum.

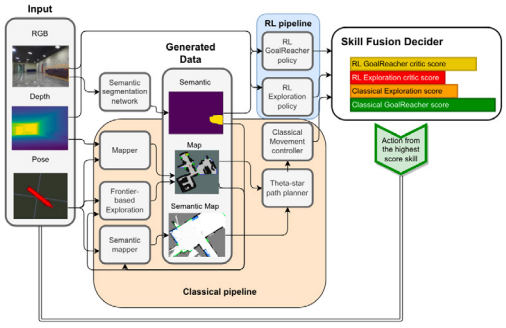
- The Graph-Enriched Biomedical Entity Representation Transformer (GEBERT) model was proposed, which uses graph neural networks based on GraphSAGE and GAT architectures to enrich language models with additional information from knowledge graphs.
In experiments on five English-language corpora, the proposed model based on the GraphSAGE architecture outperformed existing state-of-the-art named entity normalization models that do not use a knowledge graph in terms of accuracy in all cases.
The article “Graph-Enriched Biomedical Entity Representation Transformer” was published in the Proceedings of the international conference CLEF 2023 and the article “Graph-Enriched Biomedical Language Models: A Research Proposal” was published in the Proceedings of the AACL-IJCNLP SRW 2023 (CORE B).

Neural Symbolic Integration

The group led by Aleksandr Panov obtained the following results

Aleksandr Panov
Principal Research Scientist

- Several methods for solving the multi-agent pathfinding problem in a decentralized setting were proposed, relying on the integration of heuristic search and reinforcement learning methods, both model-based and model-free. One of the publications on this topic was published in the journal TNNLS (Q1, top-1 on neural network methods).
- A behavior planning module for an intelligent agent based on natural language instructions using LLM was developed. The module was tested both on specially generated data and as part of the control system of a robotic platform at the test site (together with the Sber Robotics Lab).
- The SkillFusion approach was proposed, in which classical and learnable algorithms were implemented as agent skills, and a decision-making module for selecting skills based on the assessment of each skill's internal value was developed. Based on this result, the best solution was proposed at the Habitat Challenge CVPR 2023 competition and an article was published in the Robotics journal (Q1).



Method	Success	SPL	SoftSPL
DDPPO [1]	0.18	0.10	0.35
SemExp [14]	0.24	0.14	0.26
Auxiliary RL [8]	0.51	0.29	0.34
SkillFusion	**0.64**	**0.36**	**0.38**

Skill		Metrics		
Explore	GoalReacher	Success	SPL	SoftSPL
Classical	Classical	0.410	0.182	0.263
RL	RL	0.403	0.224	0.321
RL+Classical	Classical	0.511	0.299	0.309
RL+Classical	RL+Classical	0.547	0.316	0.365

Industrial AI

The group led by Ilya Makarov obtained the following results

Ilya Makarov
Senior Researcher

In the field of Industrial ML:

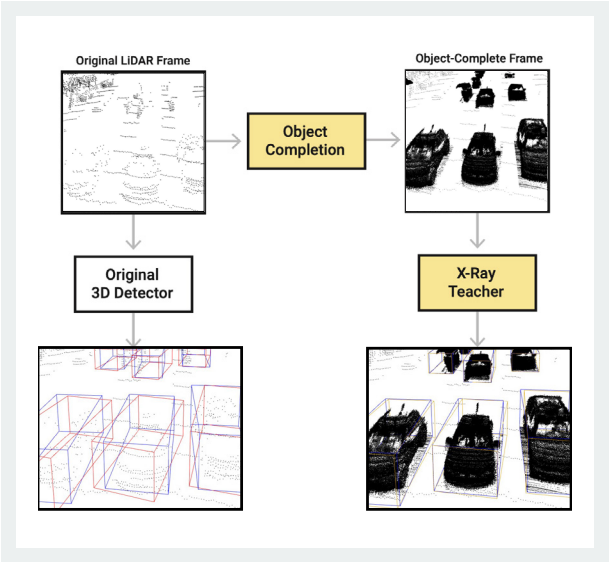
- A method for assessing the condition of equipment based on a combination of deep learning models and existing wear models, interpreted in terms of existing diagnostic rules.
- A method for organizing and interpreting low-dimensional latent space when solving the problem of industrial equipment degradation.
- A generative diffusion model of equipment failures with the construction of conditions based on trajectories of degradation close in latent space.
- A framework and large-scale experiments to protect fault diagnosis systems from adversarial attacks.
- Unsupervised Fault Diagnosis for sensor data from chemical plants:

  a. A SOTA solution was obtained based on self-supervised pre-training and deep clustering, improving quality metrics by 20–30% compared to existing approaches.

  b. It has been shown that the developed pre-training method allows models trained in a semi-supervised setting with an extremely small amount of labeled data to achieve results close to SOTA models trained in a supervised setting.

In the field of Graph ML:

- A SOTA model TI-DCGNN has been proposed for working with temporal graphs, based on proposed causality and consequence graphs, which improved the quality of short- and long-term predictions.
- A new SOTA SSL approach has been developed for pretraining models for predicting molecular properties, which preserves important domain knowledge.
- The soft graph clustering method has been adapted to the task of research in recommendation systems, which improved the quality of adaptation to changes in user behavior.



In the field of Computer Vision:

- New approaches to 3D detection based on teacher distillation on augmented data with a heavy student in supervised and semi-supervised settings in the task of automatic generation of marking. As a result:

  a. An approach based on combining LIDAR frame sequences using point cloud registration algorithms has been developed, which achieves SOTA results in the semi-supervised setting and consistently improves model quality in the supervised setting.

  b. An error has been found in the ways of comparing semi-supervised 3D detection models on the popular ONCE dataset, and a modified training method has been proposed to solve it.

In the field of Natural Language Processing:

- Organizing work on a manager assistant prototype using LLM based on a multi-agent approach (BabyAGI). The project is being carried out in collaboration with Sber AI Lab.

- A connection has been established between theoretical achievements in ontology-based data access and practical applications of machine learning, demonstrating that machine learning methods can be an effective and non-traditional approach to answering theoretical ontological queries using a covering axiom.

- An approach has been proposed that applies graph neural networks (GNN) to the task of word sense induction (WSI) using a co-occurrence network, which demonstrates satisfactory results with low computational resource consumption.

In the field of Deep Reinforcement Learning:

- Development of a model that includes advanced aggregation functions and attention mechanisms in graph convolutional networks, significantly improving the efficiency of deep reinforcement learning when navigating sparse rewards in partially observable 3D environments.

- New approaches to the task of interpretable adaptation from real to synthetic data with subsequent learning and reverse domain adaptation.

Bioinformatics

The group led by Olga Kardymon obtained the following results

Olga Kardymon
Researcher

- The GENA — LM family of language models for DNA has been developed and released into the public domain. The effectiveness of GENA for addressing a wide range of genomic tasks has been demonstrated. According to the data obtained, the GENA-LM models exhibit characteristics that match, and in some cases exceed, those of global competitors (Google, Nvidia, Stanford) presented in 2023. GENA-LM accurately predicts the locations of key genome elements: gene promoters, splicing sites, gene terminators, regulatory elements. The annotation of various genomic elements based on DNA sequence opens up possibilities for creating tools for interpreting mutations in the human genome.

- In collaboration with Dmitry Vetrov and colleagues from the Higher School of Economics (HSE), a model for unconditional diffusive generation of proteins, DIMA, has been trained on textual (amino acid) data. The algorithm has been developed for comparing the generated sequences in terms of quality and diversity, both between clusters and with competitors (EvoDiff-seq, nanoGPT, SeqDesign).

- A series of tools for studying proteins and genes have been developed and published: PROSTATA — a model for assessing protein stability upon single mutations (published in Bioinformatics, Q1, IF=5.8); a model and dataset for identifying clinically significant regions in the protein non-coding start of gene sequences (developed in collaboration with Bochkov Medical Genetics Research Center, published in Nucleic Acids Research, Q1, IF=19.16); a model for interpreting the impact of mutations in protein non-coding DNA regions on genes in different cell types (published in GigaScience, Q1, IF=7.658). The group also contributed to the creation of the book “Artificial Intelligence for Healthy Longevity,” authoring the chapter on “AI in Genomics and Epigenomics.”

Family of language models GENA-LM

Monthly downloads*	1000–3000
Downloads for the 2023*	15 883
International workshops	2

* from Hugging Face

New Materials Design

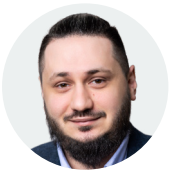
The group led by Semen Budennyy obtained the following results

Semen Budennyy
Scientific Consultant

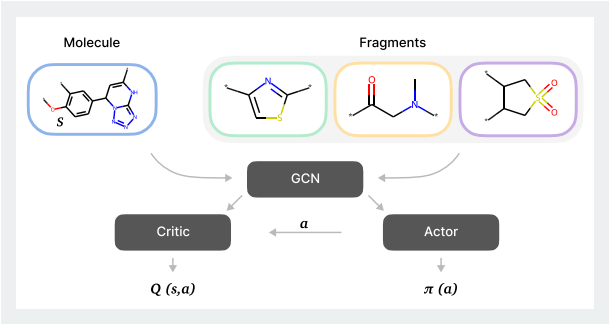
- ASCAD Light is a software suite for automatic defect detection in solar cells and localization of production line nodes related to their formation. The developed system has passed successful industrial testing at the manufacturing facility of Hevel. Scientific results have been published in Solar Energy (Q1).
- Eco4cast is an open-source Python library aimed at reducing the equivalent carbon footprint associated with AI model training by distributing cloud computing over time intervals and regions with the lowest current carbon cost of electricity. Scientific results with experimental verification of the library's efficiency have been published in Doklady Mathematics (Q2).
- ESG team solutions — ASCAD Light, Eco2AI, and Eco4cast — were presented at the UN Climate Summit.
- The "Doped $CsPbI_3$ Energetics" repository with the corresponding dataset and post-processing code. The results on predicting thermodynamic properties of Cd- and Zn-doped perovskites obtained using the developed hybrid DFT/GNN approach have been published in Computational Materials Science (Q1).
- The website "2DMD at a Glance" and repository to encourage the physical and chemical scientific community to create AI solutions in materials science. The repository implements simplified access to the open dataset of point defects in 2D materials (2DMD) and contains data processing algorithms.
- The prototype of a scientific software library for computer-aided design of 3D printable structures made from metamaterials, particularly bone implants and scaffolds.
- The surrogate model for simulating the lithographic process in microelectronics, as well as specialized metrics and loss functions that consider both the shape and topology of photo patterns. As part of a partnership with Sber Robotics Lab, the model has successfully passed the initial stage of experimental testing.

DL in Life Sciences

The group led by Artur Kadurin obtained the following results



Artur Kadurin
Research engineer



- A method for generating new molecular structures with high affinity to a target protein has been developed. The article has been published in the TMLR journal and corresponding software has been registered.
- A method for optimizing molecular conformations, which reduces computational complexity by several orders of magnitude, has been developed. The article has been accepted to the A* ICLR conference.
- New models for processing medical texts with domain knowledge represented as graphs have been developed in collaboration with Elena Tutubalina's team (accepted at the CLEF conference). Additionally, a benchmark for language models to test their chemical knowledge has been created (accepted to ICLR).
- A protein design method based on reinforcement learning is being developed in collaboration with Olga Kardymon's team.
- The second version of the benchmark and dataset for solving quantum chemical problems, nablaDFT, is being prepared for publication.
- Methods for designing flat materials are being developed in collaboration with Semen Budenny's team.
- A study was conducted in collaboration with physicists from Moscow State University to evaluate the magnetocaloric effect in doped binary materials.

NeuroInterfaces

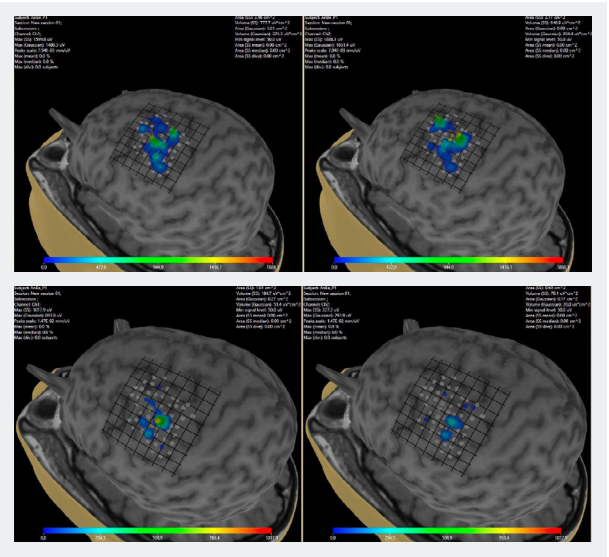
The group led by Alexey Ossadtchi obtained the following results

Alexey Ossadtchi
Leading Researcher

→ The implementation of the “Instant Neurofeedback” project results in practice.

An algorithm for instant (3 ms delay, <3% of the rhythm period) evaluation of brain rhythm phase on board the electroencephalograph NVX52 has been implemented and tested in experiments on state-dependent transcranial magnetic stimulation.

Linking the stimulation moment to the phase of the sensorimotor rhythm allowed for a several-fold increase in the accuracy of localizing cortical areas representing hand muscles. The upper row shows maps of the amplitude of the evoked motor response obtained using the standard method, while the lower row shows maps obtained in real-time using the modified NVX-52 with phase-dependent stimulation.



Learnable Intelligence

The group led by Evgeniy Burnaev obtained the following results

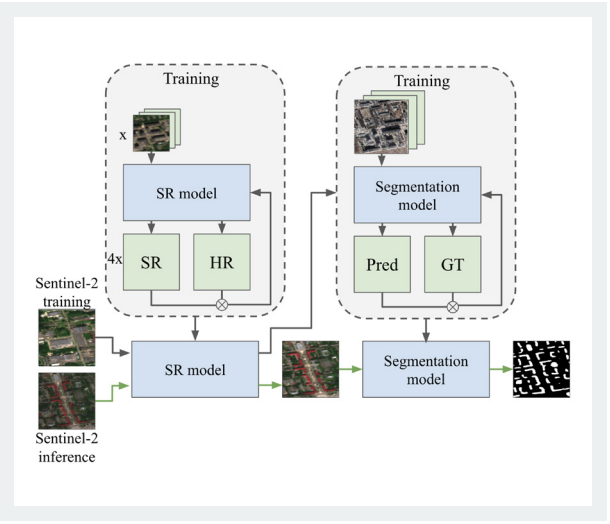
Evgeniy Burnaev
Leading Researcher

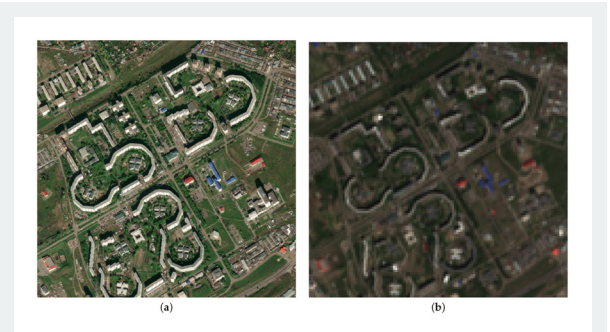
New general ML algorithms (2 Q1 papers)

- Multi-fidelity algorithm for automatic neural network architecture search
- Clustering of temporal point processes for user data processing
- Learning algorithm for a one-class support vector machine with privileged information

Applications of ML algorithms (9 Q1 papers)

- Optimal asset portfolio selection
- Detection of neurodegenerative diseases based on structural MRI and morphometric features
- Detection of depressive disorders
- Application of manifold modeling methods for analyzing human gut microbiome
- Building segmentation using remote sensing data
- Flood modeling using remote sensing data
- Land use classification using remote sensing data
- Review of ML algorithm applications in sustainable development tasks



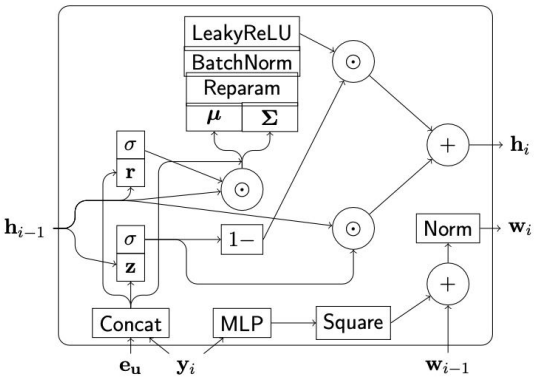


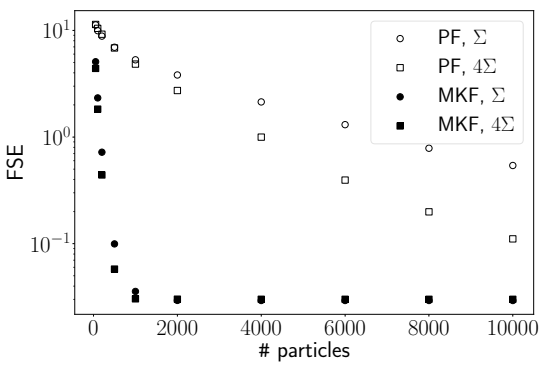
Computational Intelligence

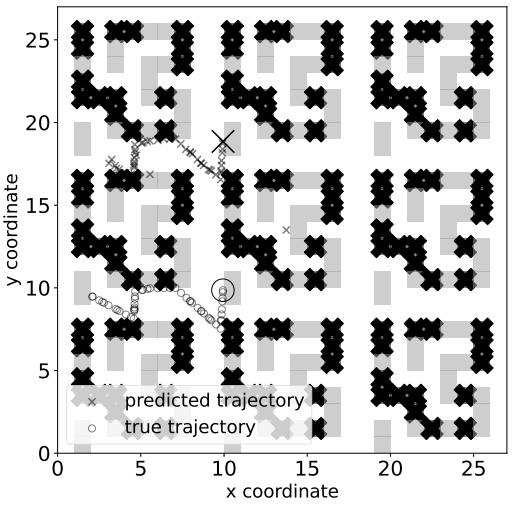
The group led by Ivan Oseledets obtained the following results

Ivan Oseledets
CEO

- A modification of the PFRNN architecture focused on object localization and significantly more efficient in terms of memory and filtering quality has been proposed.
- A method for solving the problem of object localization from noisy measurements using a combination of Kalman filter and partial filter has been developed.







Publications

A* conference papers

8 NeurIPS	3 EMNLP	5 ICLR
2 ISMAR	3 ICML	4 ACL
1 ICCV	2 CVPR	2 AAAI

NeurIPS

Neural Harmonics: Bridging Spectral Embedding and Matrix Completion in Self-Supervised Learning

Marina Munkhoeva, Ivan Oseledets

Self-supervised methods received tremendous attention thanks to their seemingly heuristic approach to learning representations that respect the semantics of the data without any apparent supervision in the form of labels. A growing body of literature is already being published in an attempt to build a coherent and theoretically grounded understanding of the workings of a zoo of losses used in modern self-supervised representation learning methods. In this paper, we attempt to provide an understanding from the perspective of a Laplace operator and connect the inductive bias stemming from the augmentation process to a low-rank matrix completion problem. To this end, we leverage the results from low-rank matrix completion to provide theoretical analysis on the convergence of modern SSL methods and a key property that affects their downstream performance. other methods on different downstream tasks.

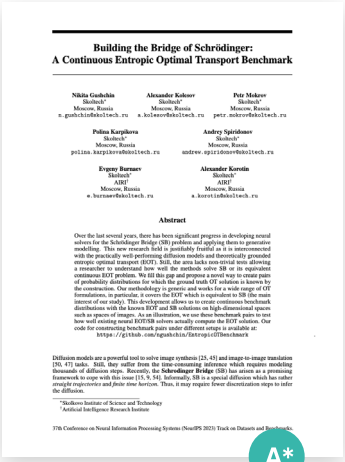
A*

Source

Building the Bridge of Schrödinger: A Continuous Entropic Optimal Transport Benchmark

Nikita Gushchin, Alexander Kolesov, Petr Mokrov, Polina Karpikova, Andrei Spiridonov, Evgeny Burnaev, Alexander Korotin

Over the last several years, there has been significant progress in developing neural solvers for the Schrödinger Bridge (SB) problem and applying them to generative modelling. This new research field is justifiably fruitful as it is interconnected with the practically well-performing diffusion models and theoretically grounded entropic optimal transport (EOT). Still, the area lacks non-trivial tests allowing a researcher to understand how well the methods solve SB or its equivalent continuous EOT problem. We fill this gap and propose a novel way to create pairs of probability distributions for which the ground truth OT solution is known by the construction. Our methodology is generic and works for a wide range of OT formulations, in particular, it covers the EOT which is equivalent to SB (the main interest of our study). This development allows us to create continuous benchmark distributions with the known EOT and SB solutions on high-dimensional spaces such as spaces of images. As an illustration, we use these benchmark pairs to test how well existing neural EOT/SB solvers actually compute the EOT solution.



A*

Source

NeurIPS

PROTES: Probabilistic Optimization with Tensor Sampling

Anastasia Batsheva, Andrei Chertkov, Gleb Ryzhakov, Ivan Oseledets

We developed a new method PROTES for black-box optimization, which is based on the probabilistic sampling from a probability density function given in the low-parametric tensor train format. We tested it on complex multidimensional arrays and discretized multivariable functions taken, among others, from real-world applications, including unconstrained binary optimization and optimal control problems, for which the possible number of elements is up to 2^{100}. In numerical experiments, both on analytic model functions and on complex problems, PROTES outperforms existing popular discrete optimization methods (Particle Swarm Optimization, Covariance Matrix Adaptation, Differential Evolution, and others).


PROTES: Probabilistic Optimization with Tensor Sampling

Anastasia Batsheva* Skolkovo Institute of Science and Technology Moscow, Russia a.batsheva@skoltech.ru

Andrei Chertkov* Skolkovo Institute of Science and Technology Moscow, Russia a.chertkov@skoltech.ru

Gleb Ryzhakov* Skolkovo Institute of Science and Technology Moscow, Russia g.ryzhakov@skoltech.ru

Ivan Oseledets Skolkovo Institute of Science and Technology and AIRI, Moscow, Russia i.oseledets@skoltech.ru

Abstract

We developed a new method PROTES for black-box optimization, which is based on the probabilistic sampling from a probability density function given in the low-parametric tensor train format. We tested it on complex multidimensional arrays and discretized multivariable functions taken, among others, from real-world applications, including unconstrained binary optimization and optimal control problems, for which the possible number of elements is up to 2^{100}. In numerical experiments, both on analytic model functions and on complex problems, PROTES outperforms existing popular discrete optimization methods (Particle Swarm Optimization, Covariance Matrix Adaptation, Differential Evolution, and others).

1 Introduction

The multidimensional optimization problem is one of the most common in machine learning. It includes the important case of gradient-free discrete optimization [30, 20, 21, 41]

$$x_{min} = \min_{\mathbf{x}} f(\mathbf{x}), \quad x_* \equiv [x_1, x_2, \ldots, x_d], \quad x_k \in \{1, 2, \ldots, N_k\}, \tag{1}$$

where d is the dimensionality of the problem, and $N_1, N_2, \ldots, N_d$ are the numbers of items for each dimension. Such settings arise when searching for the minimum or maximum element in an implicitly given multidimensional array (tensor), including when considering the discretization of functions from a continuous argument. Multidimensional discrete optimization problems are still computationally difficult in the case of complex target functions or large dimensions, and efficient direct gradient-free optimization procedures are highly needed.

The development of methods for low-rank tensor approximations has made it possible to introduce fundamentally new approaches for the approximation, storage, and operation with multidimensional

*Equal contribution.

A tensor is just a multidimensional array with several dimensions d ($d \geq 1$). A two-dimensional tensor ($d = 2$) is a matrix, and when $d = 1$ it is a vector. For scalars we use normal font ($a, b, c, \ldots$), we denote vectors with bold letters ($\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$), we use upper case letters ($A, B, C, \ldots$) for matrices, and calligraphic upper case letters ($\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$) for tensors with $d \geq 3$. The $(i_1, i_2, \ldots, i_d)$-th entry of a d-dimensional tensor $\mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times \ldots \times n_d}$ is denoted by $y = \mathcal{Y}[i_1, i_2, \ldots, i_d]$, where $i_k = 1, 2, \ldots, N_k$ ($k = 1, 2, \ldots, d$), and N_k is a size of the k-th mode. The mode-i slice of such tensor is denoted by $\mathcal{Y}[i_1, \ldots, i_{k-1}, :, i_{k+1}, \ldots, i_d]$.

Preprint. Under review.

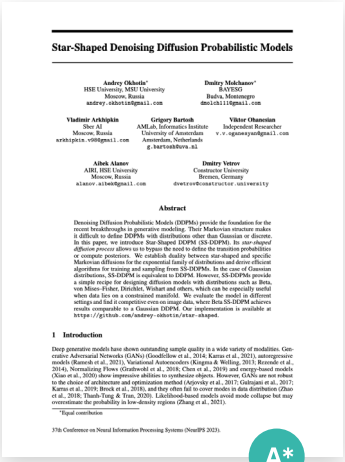
A*

Source

Star-Shaped Denoising Diffusion Probabilistic Models

Andrey Okhotin, Dmitry Molchanov, Vladimir Arkhipkin, Grigory Bartosh, Viktor Ohanesian, Aibek Alanov, Dmitry Vetrov

Denoising Diffusion Probabilistic Models (DDPMs) provide the foundation for the recent breakthroughs in generative modeling. Their Markovian structure makes it difficult to define DDPMs with distributions other than Gaussian or discrete. In this paper, we introduce Star-Shaped DDPM (SS-DDPM). Its star-shaped diffusion process allows us to bypass the need to define the transition probabilities or compute posteriors. We establish duality between star-shaped and specific Markovian diffusions for the exponential family of distributions and derive efficient algorithms for training and sampling from SS-DDPMs. In the case of Gaussian distributions, SS-DDPM is equivalent to DDPM. However, SS-DDPMs provide a simple recipe for designing diffusion models with distributions such as Beta, von Mises–Fisher, Dirichlet, Wishart and others, which can be especially useful when data lies on a constrained manifold. We evaluate the model in different settings and find it competitive even on image data, where Beta SS-DDPM achieves results comparable to a Gaussian DDPM.


Star-Shaped Denoising Diffusion Probabilistic Models

Andrey Okhotin* HSE University, MIPT University Moscow, Russia andrey.okhotin@gmail.com

Dmitry Molchanov* BAYESG Budva, Montenegro dmolchi11@gmail.com

Vladimir Arkhipkin Sber AI Moscow, Russia arkhipkin.v98@gmail.com

Grigory Bartosh AMLab, Informatics Institute University of Amsterdam Amsterdam, Netherlands g.bartosh@uva.nl

Viktor Ohanesian Independent Researcher v.v.oganesyan@gmail.com

Aibek Alanov AIRI, HSE University Moscow, Russia alanov.aibek@gmail.com

Dmitry Vetrov Constructor University Bremen, Germany dvetrov@constructor.university

Abstract

Denoising Diffusion Probabilistic Models (DDPMs) provide the foundation for the recent breakthroughs in generative modeling. Their Markovian structure makes it difficult to define DDPMs with distributions other than Gaussian or discrete. In this paper, we introduce Star-Shaped DDPM (SS-DDPM). Its *star-shaped diffusion process* allows us to bypass the need to define the transition probabilities or compute posteriors. We establish duality between star-shaped and specific Markovian diffusions for the exponential family of distributions and derive efficient algorithms for training and sampling from SS-DDPMs. In the case of Gaussian distributions, SS-DDPM is equivalent to DDPM. However, SS-DDPMs provide a simple recipe for designing diffusion models with distributions such as Beta, von Mises–Fisher, Dirichlet, Wishart and others, which can be especially useful when data lies on a constrained manifold. We evaluate the model in different settings and find it competitive even on image data, where Beta SS-DDPM achieves results comparable to a Gaussian DDPM. Our implementation is available at https://github.com/andrey-olkhotin/star-shaped.

1 Introduction

Deep generative models have shown outstanding sample quality in a wide variety of modalities. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Karras et al., 2021), autoregressive models (Ramesh et al., 2021), Variational Autoencoders (Kingma & Welling, 2013; Rezende et al., 2014), Normalizing Flows (Grathwohl et al., 2018; Chen et al., 2019) and energy-based models (Xiao et al., 2020) show impressive abilities to synthesize objects. However, GANs are not robust to the choice of architecture and optimization method (Arjovsky et al., 2017; Gulrajani et al., 2017; Karras et al., 2019; Brock et al., 2018), and they often fail to cover modes in data distribution (Zhao et al., 2018; Thanh-Tung & Tran, 2020). Likelihood-based models avoid mode collapse but may overestimate the probability in low-density regions (Zhang et al., 2021).

*Equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A*

Source

NeurIPS

Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts

Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, Evgeny Burnaev

Rapidly increasing quality of AI-generated content makes it difficult to distinguish between human and AI-generated texts, which may lead to undesirable consequences for society. Therefore, it becomes increasingly important to study the properties of human texts that are invariant over different text domains and varying proficiency of human writers, can be easily calculated for any language, and can robustly separate natural and AI-generated texts regardless of the generation model and sampling method. In this work, we propose such an invariant for human-written texts, namely the intrinsic dimensionality of the manifold underlying the set of embeddings for a given text sample. We show that the average intrinsic dimensionality of fluent texts in a natural language is hovering around the value 9 for several alphabet-based languages and around 7 for Chinese, while the average intrinsic dimensionality of AI-generated texts for each language is ≈1.5 lower, with a clear statistical separation between human-generated and AI-generated distributions. This property allows us to build a score-based artificial text detector. The proposed detector's accuracy is stable over text domains, generator models, and human...


Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts

Abstract

1 Introduction

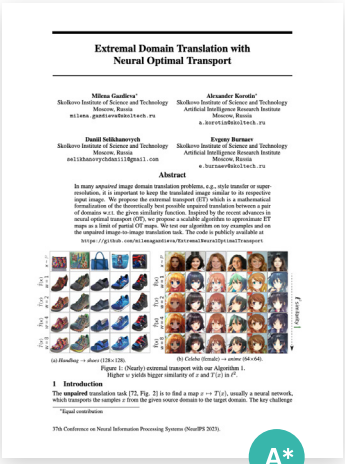
A*

Source

Extremal Domain Translation with Neural Optimal Transport

Milena Gazdieva, Alexander Korotin, Daniil Selikhanovych, Evgeny Burnaev

In many unpaired image domain translation problems, e.g., style transfer or super-resolution, it is important to keep the translated image similar to its respective input image. We propose the extremal transport (ET) which is a mathematical formalization of the theoretically best possible unpaired translation between a pair of domains w.r.t. the given similarity function. Inspired by the recent advances in neural optimal transport (OT), we propose a scalable algorithm to approximate ET maps as a limit of partial OT maps. We test our algorithm on toy examples and on the unpaired image-to-image translation task.


Extremal Domain Translation with Neural Optimal Transport

Abstract

1 Introduction

A*

Source

NeurIPS

To Stay or Not to Stay in the Pre-train Basin: Insights on Ensembling in Transfer Learning

Ildus Sadrtdinov, Dmitrii Pozdeev, Dmitry Vetrov, Ekaterina Lobacheva

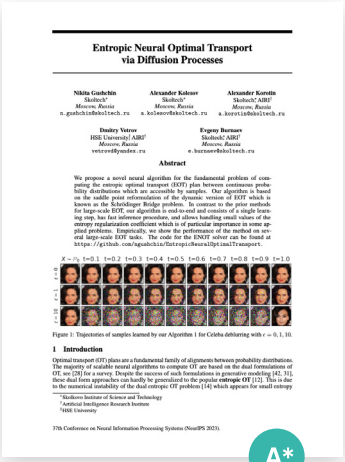
Transfer learning and ensembling are two popular techniques for improving the performance and robustness of neural networks. Due to the high cost of pre-training, ensembles of models fine-tuned from a single pre-trained checkpoint are often used in practice. Such models end up in the same basin of the loss landscape, which we call the pre-train basin, and thus have limited diversity. In this work, we show that ensembles trained from a single pre-trained checkpoint may be improved by better exploring the pre-train basin, however, leaving the basin results in losing the benefits of transfer learning and in degradation of the ensemble quality. Based on the analysis of existing exploration methods, we propose a more effective modification of the Snapshot Ensembles (SSE) for transfer learning setup, StarSSE, which results in stronger ensembles and uniform model soups.

Source

Entropic Neural Optimal Transport via Diffusion Processes

Nikita Gushchin, Alexander Kolesov, Alexander Korotin, Dmitry Vetrov, Evgeny Burnaev

We propose a novel neural algorithm for the fundamental problem of computing the entropic optimal transport (EOT) plan between continuous probability distributions which are accessible by samples. Our algorithm is based on the saddle point reformulation of the dynamic version of EOT which is known as the Schrödinger Bridge problem. In contrast to the prior methods for large-scale EOT, our algorithm is end-to-end and consists of a single learning step, has fast inference procedure, and allows handling small values of the entropy regularization coefficient which is of particular importance in some applied problems. Empirically, we show the performance of the method on several large-scale EOT tasks.



Source

EMNLP

Kandinsky: an Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, Denis Dimitrov

Text-to-image generation is a significant domain in modern computer vision and has achieved substantial improvements through the evolution of generative architectures. Among these, there are diffusion-based models that have demonstrated essential quality enhancements. These models are generally split into two categories: pixel-level and latent-level approaches. We present Kandinsky, a novel exploration of latent diffusion architecture, combining the principles of the image prior models with latent diffusion techniques. The image prior model is trained separately to map text embeddings to image embeddings of CLIP. Another distinct feature of the proposed model is the modified MoVQ implementation, which serves as the image autoencoder component. Overall, the designed model contains 3.3B parameters. We also deployed a user-friendly demo system that supports diverse generative modes such as text-to-image generation, image fusion, text and image fusion, image variations generation, and text-guided inpainting/outpainting. Additionally, we released the source code and...

Kandinsky: an Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion

A*

Source

Better Together: Enhancing Generative Knowledge Graph Completion with Language Models and Neighborhood Information

Alla Chepurova, Aydar Bulatov, Yuri Kuratov, Mikhail S. Burtsev

Real-world Knowledge Graphs (KGs) often suffer from incompleteness, which limits their potential performance. Knowledge Graph Completion (KGC) techniques aim to address this issue. However, traditional KGC methods are computationally intensive and impractical for large-scale KGs, necessitating the learning of dense node embeddings and computing pairwise distances. Generative transformer-based language models (e.g., T5 and recent KGT5) offer a promising solution as they can predict the tail nodes directly. In this study, we propose to include node neighborhoods as additional information to improve KGC methods based on language models. We examine the effects of this imputation and show that, on both inductive and transductive Wikidata subsets, our method outperforms KGT5 and conventional KGC approaches...

Better Together: Enhancing Generative Knowledge Graph Completion with Language Models and Neighborhood Information

A*

Source

EMNLP

LM-Polygraph: Uncertainty Estimation for Language Models

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Artem Shelmanov

Recent advancements in the capabilities of large language models (LLMs) have paved the way for a myriad of groundbreaking applications in various fields. However, a significant challenge arises as these models often «hallucinate», i.e., fabricate facts without providing users an apparent means to discern the veracity of their statements. Uncertainty estimation (UE) methods are one path to safer, more responsible, and more effective use of LLMs. However, to date, research on UE methods for LLMs has been focused primarily on theoretical rather than engineering contributions. In this work, we tackle this issue by introducing LM-Polygraph, a framework with implementations of a battery of state-of-the-art UE methods for LLMs in text generation tasks, with unified program interfaces in Python. Additionally, it introduces an extendable benchmark for consistent evaluation of UE techniques by researchers, and a demo web application that enriches the standard chat dialog with confidence scores, empowering end-users to discern unreliable responses. LM-Polygraph is compatible with the most recent LLMs, including BLOOMz, LLaMA-2, ChatGPT, and GPT-4, and is designed to support future releases of similarly-styled LMs.


LM-Polygraph: Uncertainty Estimation for Language Models

Abstract

1 Introduction

A*

Source

ICLR

Understanding DDPM Latent Codes Through Optimal Transport

Valentin Khrulkov, Gleb Ryzhakov, Andrei Chertkov, Ivan Oseledets

Diffusion models have recently outperformed alternative approaches to model the distribution of natural images. Such diffusion models allow for deterministic sampling via the probability flow ODE, giving rise to a latent space and an encoder map. While having important practical applications, such as the estimation of the likelihood, the theoretical properties of this map are not yet fully understood. In the present work, we partially address this question for the popular case of the VP-SDE (DDPM) approach. We show that, perhaps surprisingly, the DDPM encoder map coincides with the optimal transport map for common distributions; we support this claim by extensive numerical experiments using advanced tensor train solver for multidimensional Fokker-Planck equation. We provide additional theoretical evidence for the case of multivariate normal distributions.

Source

Constructive TT-representation of the tensors given as index interaction functions with applications

Gleb V. Ryzhakov, Ivan V. Oseledets

This paper presents a method to build explicit tensor-train (TT) representations. We show that a wide class of tensors can be explicitly represented with sparse TT-cores, obtaining, in many cases, optimal TT-ranks. Numerical experiments show that our method outperforms the existing ones in several practical applications, including game theory problems. Theoretical estimations of the number of operations show that in some problems, such as permanent calculation, our methods are close to the known optimal asymptotics, which are obtained by a completely different type of methods.

Source

ICLR

Learning topology-preserving data representations

Ilya Trofimov, Daniil Cherniavskii, Eduard Tulchinskii, Nikita Balabin, Evgeny Burnaev, Serguei Barannikov

We propose a method for learning topology-preserving data representations (dimensionality reduction). The method aims to provide topological similarity between the data manifold and its latent representation via enforcing the similarity in topological features (clusters, loops, 2D voids, etc.) and their localization. The core of the method is the minimization of the Representation Topology Divergence (RTD) between original high-dimensional data and low-dimensional representation in latent space. RTD minimization provides closeness in topological features with strong theoretical guarantees. We develop a scheme for RTD differentiation and apply it as a loss term for the autoencoder. The proposed method «RTD-AE» better preserves the global structure and topology of the data manifold than state-of-the-art competitors as measured by linear correlation, triplet distance ranking accuracy, and Wasserstein distance between persistence barcodes.

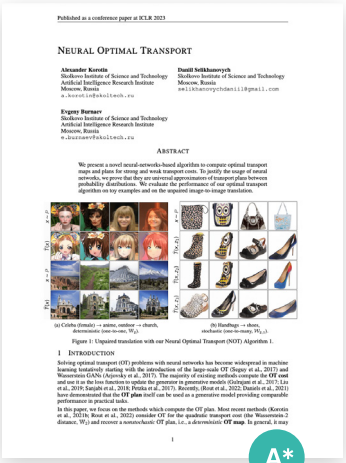
A*

Source

Neural Optimal Transport

Alexander Korotin, Daniil Selikhanovych, Evgeny Burnaev

We present a novel neural-networks-based algorithm to compute optimal transport maps and plans for strong and weak transport costs. To justify the usage of neural networks, we prove that they are universal approximators of transport plans between probability distributions. We evaluate the performance of our optimal transport algorithm on toy examples and on the unpaired image-to-image translation.



A*

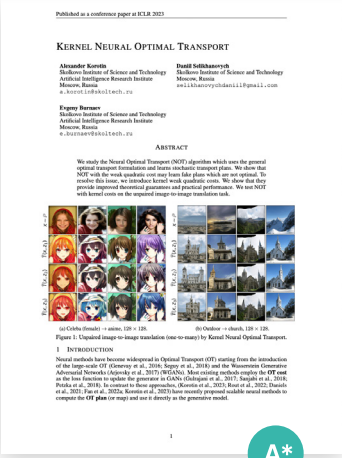
Source

ICLR

Kernel Neural Optimal Transport

Alexander Korotin, Daniil Selikhanovych, Evgeny Burnaev

We study the Neural Optimal Transport (NOT) algorithm which uses the general optimal transport formulation and learns stochastic transport plans. We show that NOT with the weak quadratic cost may learn fake plans which are not optimal. To resolve this issue, we introduce kernel weak quadratic costs. We show that they provide improved theoretical guarantees and practical performance. We test NOT with kernel costs on the unpaired image-to-image translation task.



Published as a conference paper at ICLR 2023

KERNEL NEURAL OPTIMAL TRANSPORT

Alexander Korotin
Skolkovo Institute of Science and Technology
Artificial Intelligence Research Institute
Moscow, Russia
a.korotin@skoltech.ru

Daniil Selikhanovych
Skolkovo Institute of Science and Technology
Moscow, Russia
selikhanovychdaniil@gmail.com

Evgeny Burnaev
Skolkovo Institute of Science and Technology
Artificial Intelligence Research Institute
Moscow, Russia
e.burnaev@skoltech.ru

ABSTRACT

We study the Neural Optimal Transport (NOT) algorithm which uses the general optimal transport formulation and learns stochastic transport plans. We show that NOT with the weak quadratic cost may learn fake plans which are not optimal. To resolve this issue, we introduce kernel weak quadratic costs. We show that they provide improved theoretical guarantees and practical performance. We test NOT with kernel costs on the unpaired image-to-image translation task.

(a) Celeba (female) → anime, 128 × 128.

(b) Outdoor → church, 128 × 128.

Figure 1: Unpaired image-to-image translation (one-to-many) by Kernel Neural Optimal Transport.

1 INTRODUCTION

Neural methods have become widespread in Optimal Transport (OT) starting from the introduction of the large-scale OT (Genevay et al., 2016; Seguy et al., 2018) and the Wasserstein Generative Adversarial Networks (Arjovsky et al., 2017) (WGANs). Most existing methods employ the **OT cost** as the loss function to update the generator in GANs (Gulrajani et al., 2017; Sanjabi et al., 2018; Petzka et al., 2018). In contrast to these approaches, (Korotin et al., 2023; Rout et al., 2022; Daniels et al., 2021; Fan et al., 2022a; Korotin et al., 2021) have recently proposed scalable neural methods to compute the **OT plan** (or map) and use it directly as the generative model.

1

A*

Source

ISMAR

SwiftDepth: An Efficient Hybrid CNN-Transformer Model for Self-Supervised Monocular Depth Estimation on Mobile Devices

Albert Luginov, Ilya Makarov

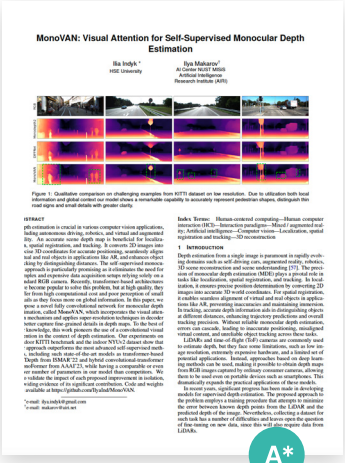
A*

Source

Self-supervised Monocular Depth Estimation (MDE) models trained solely on single-camera video have gained significant popularity. Recent studies have shown that Vision Transformers (ViT) can improve depth estimation quality, despite their high computational demands. In Extended Reality (XR) context, lightweight and fast models are crucial for seamless operation on mobile devices. This paper proposes SwiftDepth, a hybrid MDE framework that fulfils these requirements. The model combines the benefits of Convolutional Neural Network (CNN), which provides speed and shift invariance, and ViT, which offers a global receptive field. We utilize SwiftFormer, a low-latency feature extraction network with efficient additive attention. Also, we introduce a novel two-level decoder to enhance depth estimation quality without an increase in the number of parameters. Our model achieves comparable results to the state-of-the-art lightweight Lite-Mono on the KITTI...

MonoVAN: Visual Attention for Self-Supervised Monocular Depth Estimation

Iia Indyk, Ilya Makarov



A*

Source

Depth estimation is crucial in various computer vision applications, including autonomous driving, robotics, and virtual and augmented reality. An accurate scene depth map is beneficial for localization, spatial registration, and tracking. It converts 2D images into precise 3D coordinates for accurate positioning, seamlessly aligns virtual and real objects in applications like AR, and enhances object tracking by distinguishing distances. The self-supervised monocular approach is particularly promising as it eliminates the need for complex and expensive data acquisition setups relying solely on a standard RGB camera. Recently, transformer-based architectures have become popular to solve this problem, but at high quality, they suffer from high computational cost and poor perception of small details as they focus more on global information. In this paper, we propose a novel fully convolutional network for monocular depth estimation, called MonoVAN, which incorporates the visual attention mechanism and...

ICML

Fast Rates for Maximum Entropy Exploration

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello,
Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault,
Yunhao Tang, Michal Valko, Pierre Menard

We address the challenge of exploration in reinforcement learning (RL) when the agent operates in an unknown environment with sparse or no rewards. In this work, we study the maximum entropy exploration problem of two different types. The first type is visitation entropy maximization previously considered by Hazan et al. (2019) in the discounted setting. For this type of exploration, we propose a game-theoretic algorithm that has $\tilde{O}(H^3S^2A/\varepsilon^2)$ sample complexity thus improving the ε-dependence upon existing results, where S is a number of states, A is a number of actions, H is an episode length, and ε is a desired accuracy. The second type of entropy we study is the trajectory entropy. This objective function is closely related to the entropy-regularized MDPs, and we propose a simple algorithm that has a sample complexity of order $\tilde{O}(poly(S, A, H)/\varepsilon)$. Interestingly, it is the first theoretical result in RL literature that establishes the potential statistical advantage of regularized MDPs for exploration. Finally, we apply developed regularization techniques to reduce sample complexity of visitation entropy maximization to $\tilde{O}(H^2SA/\varepsilon^2)$, yielding a statistical separation between maximum entropy exploration and reward-free exploration.

Source

ICML

General Covariance Data Augmentation for Neural PDE Solvers

Vladimir Fanaskov, Tianchi Yu, Alexander Rudikov, Ivan V. Oseledets

The growing body of research shows how to replace classical partial differential equation (PDE) integrators with neural networks. The popular strategy is to generate the input-output pairs with a PDE solver, train the neural network in the regression setting, and use the trained model as a cheap surrogate for the solver. The bottleneck in this scheme is the number of expensive queries of a PDE solver needed to generate the dataset. To alleviate the problem, we propose a computationally cheap augmentation strategy based on general covariance and simple random coordinate transformations. Our approach relies on the fact that physical laws are independent of the coordinate choice, so the change in the coordinate system preserves the type of a parametric PDE and only changes PDE's data (e.g., initial conditions, diffusion coefficient). For tried neural networks and partial differential equations, proposed augmentation improves test error by 23% on average. The worst observed result is a 17% increase in test error for multilayer perceptron, and the best case is a 80% decrease for dilated residual network.


General Covariance Data Augmentation for Neural PDE Solvers

A*

Source

Few-bit Backward: Quantized Gradients of Activation Functions for Memory Footprint Reduction

Georgii Novikov, Daniel Bershatsky, Julia Gusak, Alex Shonenkov, Denis Dimitrov, Ivan Oseledets

Memory footprint is one of the main limiting factors for large neural network training. In backpropagation, one needs to store the input to each operation in the computational graph. Every modern neural network model has quite a few pointwise nonlinearities in its architecture, and such operations induce additional memory costs that, as we show, can be significantly reduced by quantization of the gradients. We propose a systematic approach to compute optimal quantization of the retained gradients of the pointwise nonlinear functions with only a few bits per each element. We show that such approximation can be achieved by computing an optimal piecewise-constant approximation of the derivative of the activation function, which can be done by dynamic programming. The drop-in replacements are implemented for all popular nonlinearities and can be used in any existing...


Few-bit Backward: Quantized Gradients of Activation Functions for Memory Footprint Reduction

A*

Source

ACL

Efficient Out-of-Domain Detection for Sequence to Sequence Models

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, Artem Shelmanov

Sequence-to-sequence (seq2seq) models based on the Transformer architecture have become a ubiquitous tool applicable not only to classical text generation tasks such as machine translation and summarization but also to any other task where an answer can be represented in a form of a finite text fragment (e.g., question answering). However, when deploying a model in practice, we need not only high performance but also an ability to determine cases where the model is not applicable. Uncertainty estimation (UE) techniques provide a tool for identifying out-of-domain (OOD) input where the model is susceptible to errors. State-of-the-art UE methods for seq2seq models rely on computationally heavyweight and impractical deep ensembles. In this work, we perform an empirical investigation of various novel UE methods for large pre-trained seq2seq models T5 and BART on three tasks: machine translation, text summarization, and question answering. We apply computationally lightweight density-based UE methods to seq2seq models and show that they often outperform heavyweight deep ensembles on the task of OOD detection.


Efficient Out-of-Domain Detection for Sequence to Sequence Models

Artem Vazhentsev[1,2,3], Akim Tsvigun[4,5,2], Roman Vashurin[4,3], Sergey Petrakov[3], Daniil Vasilev[4], Maxim Panov[4], Alexander Panchenko[2,3], and Artem Shelmanov[3]

[1]AIRI, [2]Skoltech, [3]MBZUAI, [4]ITI, [5]HSE University, [6]AI Center NUST MISiS, [7]Sentrush

{vazhentsev, panchenko}@airi.net {sergey.petrakov}@skoltech.ru

{roman.vashurin, maxim.panov}@mbzuai.ac.ae artem.shelmanov@mbzuai.ac.ae

Abstract

Sequence-to-sequence (seq2seq) models based on the Transformer architecture have become a ubiquitous tool applicable not only to classical text generation tasks such as machine translation and summarization but also to any other task where an answer can be represented in a form of a finite text fragment (e.g., question answering). However, when deploying a model in practice, we need not only high performance but also an ability to determine cases where the model is not applicable. Uncertainty estimation (UE) techniques provide a tool for identifying out-of-domain (OOD) input where the model is susceptible to errors. State-of-the-art UE methods for seq2seq models rely on computationally heavyweight and impractical deep ensembles. In this work, we perform an empirical investigation of various novel UE methods for large pre-trained seq2seq models T5 and BART on three tasks: machine translation, text summarization, and question answering. We apply computationally lightweight density-based UE methods to seq2seq models and show that they often outperform heavyweight deep ensembles on the task of OOD detection[1].

1 Introduction

Sequence-to-sequence (seq2seq) models achieve state-of-the-art performance in various NLP tasks, such as neural machine translation (NMT, Vaswani et al. (2017); Song et al. (2019), Zhu et al. (2020); Liu et al. (2020)), abstractive text summarization (ATS; Zhang et al. (2020); Lewis et al. (2020)), question answering (QA; Raffel et al. (2020)), and others. Such models may encounter various user inputs when exposed to the general public. In many cases, it is preferable to detect and handle in a special way what is known as out-of-domain (OOD) inputs. OOD instances are significantly different

[1]The code for reproducing experiments is available online at https://github.com/skat-nlp/seq2seq_ood_detection

∗ Equal contribution

from the data used during training, and as a result, model predictions on such inputs might be unreliable. OOD can be performed in supervised and unsupervised ways. In a supervised approach, one trains a discriminator between in-domain (ID) and OOD instances on a labeled dataset of such instances, which is manually annotated (Hendrycks et al., 2019) or synthetically generated (Liang et al., 2018). The drawback of such an approach is that the discriminator is also limited in what instances it can correctly process. Therefore, in many practical cases, it might be better to use an unsupervised approach, where OOD instances are detected using uncertainty estimation (UE) methods.

Related work. UE for text generation models is still an area of ongoing research with only a limited number of works. Malinin and Gales (2020) propose various ensemble-based UE methods for seq2seq models and evaluate them on two tasks: NMT and automatic speech recognition. Ensemble-based methods in conjunction with Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) are also investigated in (Lakovskaya et al., 2021). The authors find that the ensemble-based UE methods lead to the best results for OOD detection in the neural semantic parsing task. Xiao et al. (2020) introduce a novel UE method BLEUVar, which is also based on MC dropout. The uncertainty score is calculated as a sum of the squared complements of BLEU scores for all pairs of generated texts obtained with different dropout masks. The method shows improvements over the baselines in NMT. Lyo et al. (2020) further explore this method for OOD detection in question answering. Gidiotis and Tsoumakas (2022) show that BLEUVar can also be applied for UE in summarization. The aforementioned methods entail performing multiple model inferences for each individual input, resulting in high computational overhead. Recently, Kuhn et al. (2022) propose a method that does not leverage MC dropout, but samples multiple predictions without

1430

Findings of the Association for Computational Linguistics: ACL 2023, pages 1430–1...
July 9-14, 2023 ©2023 Association for Computational Linguistics

A*

Source

Layerwise universal adversarial attack on NLP models

O Tsymboi, D Malaev, A Petrovskii, I Oseledets

In this work, we examine the vulnerability of language models to universal adversarial triggers (UATs). We propose a new white-box approach to the construction of layerwise UATs (LUATs), which searches the triggers by perturbing hidden layers of a network. On the example of three transformer models and three datasets from the GLUE benchmark, we demonstrate that our method provides better transferability in a model-to-model setting with an average gain of 9.3% in the fooling rate over the baseline. Moreover, we investigate triggers transferability in the task-to-task setting. Using small subsets from the datasets similar to the target tasks for choosing a perturbed layer, we show that LUATs are more efficient than vanilla UATs by 7.1% in the fooling rate.


Layerwise universal adversarial attack on NLP models

Otga Tsymboi[1,2], Daniil Malaev[1], Andrei Petrovskii[1], and Ivan Oseledets[3,4]

[1]Sber AI Lab, Moscow, Russia

[2]Moscow Institute of Physics and Technology, Moscow, Russia

[3]Skolkovo Institute of Science and Technology, Moscow, Russia

[4]Artificial Intelligence Research Institute (AIRI), Moscow, Russia

tsimboy.oa@phystech.edu, n925517132@gmail.com, andreypetrovskij@gmail.com

i.oseledets@skoltech.ru

Abstract

In this work, we examine the vulnerability of language models to universal adversarial triggers (UATs). We propose a new white-box approach to the construction of layerwise UATs (LUATs), which searches the triggers by perturbing hidden layers of a network. On the example of three transformer models and three datasets from the GLUE benchmark, we demonstrate that our method provides better transferability in a model-to-model setting with an average gain of 9.3% in the fooling rate over the baseline. Moreover, we investigate triggers transferability in the task-to-task setting. Using small subsets from the datasets similar to the target tasks for choosing a perturbed layer, we show that LUATs are more efficient than vanilla UATs by 7.1% in the fooling rate.

1 Introduction

One of the fundamental drawbacks of modern neural networks is their vulnerability to adversarial attacks (Szegedy et al., 2013; Goodfellow et al., 2014), imperceptible perturbations to the data samples that leave the ground truth label unchanged but are able to modify model prediction drastically. The samples obtained as the result of these perturbations are called adversarial examples. First discovered for image datasets (Szegedy et al., 2013), this phenomenon was then demonstrated for other types of data, including natural language (Papernot et al., 2016; Liang et al., 2017; Gao et al., 2018).

The originally proposed methods of adversarial attack construction were sample-dependent, which means that one can not apply the same perturbation to different dataset items and expect equal success. Sample-agnostic, or in other words, universal, adversarial perturbations (UAPs) were proposed in Moosavi-Dezfooli et al. (2017), where based on a small subset of image data, the authors constructed perturbations leading to prediction change of 80-90% (depending on the model) of samples. They also showed that UAPs discovered on one model could successfully fool another.

The generalization of the universal attacks to natural language data was made by Wallace et al. (2019). Short additives (triggers) were inserted at the beginning of data samples, and then a search over the token space, in order to maximize the probability of the negative class on the chosen data subset, was performed. The found triggers turned out to be very efficient at fooling the model, repeating the success of UAPs proposed for images.

The conventional way to look for adversarial examples is to perturb the output of a model. Considering image classification neural networks, Khrulkov and Oseledets (2018) proposed to search for perturbations to hidden layers by approximating the so-called (p, q)-singular vectors (Boyd, 1974) of the corresponding Jacobian matrix. Then one can hope that the error will propagate through the whole network end, resulting in model prediction change. They showed that this approach allows obtaining a high fooling rate based on significantly smaller data subsets than those leveraged by Moosavi-Dezfooli et al. (2017).

In this paper, we aim to continue the investigation of neural networks' vulnerability to universal adversarial attacks in the case of natural language data. Inspired by the approach considered by Khrulkov and Oseledets (2018), we look for the perturbations to hidden layers of a model instead of the loss function. In order to avoid projection from embedding to discrete space, we use simplex parametrization of the search space (see, e.g. (Dong et al., 2021; Gao et al., 2021)). We formulate the corresponding optimization problem and propose the algorithm for obtaining its approximate solution. On the example of three transformer models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020) and three GLUE datasets (Wang et al., 2018) we demonstrate higher efficiency of our method over the original approach

129

Findings of the Association for Computational Linguistics: ACL 2023, pages 128–141...
July 9-14, 2023 ©2023 Association for Computational Linguistics

A*

Source

ACL

Hybrid Uncertainty Quantification for Selective Text Classification in Ambiguous Tasks

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, Artem Shelmanov

Many text classification tasks are inherently ambiguous, which results in automatic systems having a high risk of making mistakes, in spite of using advanced machine learning models. For example, toxicity detection in user-generated content is a subjective task, and notions of toxicity can be annotated according to a variety of definitions that can be in conflict with one another. Instead of relying solely on automatic solutions, moderation of the most difficult and ambiguous cases can be delegated to human workers. Potential mistakes in automated classification can be identified by using uncertainty estimation (UE) techniques. Although UE is a rapidly growing field within natural language processing, we find that state-of-the-art UE methods estimate only epistemic uncertainty and show poor performance, or under-perform trivial methods for ambiguous tasks such as toxicity detection. We argue that in order to create robust uncertainty estimation methods for ambiguous tasks it is necessary to account also for aleatoric uncertainty. In this paper, we propose a new uncertainty...

Source

A System for Answering Simple Questions in Multiple Languages

A Razzhigaev, M Salnikov, V Malykh, P Braslavski, A Panchenko

Our research focuses on the most prevalent type of queries – simple questions — exemplified by questions like “What is the capital of France?”. These questions reference an entity such as “France”, which is directly connected (one hop) to the answer entity “Paris” in the underlying knowledge graph (KG). We propose a multilingual Knowledge Graph Question Answering (KGQA) technique that orders potential responses based on the distance between the question’s text embeddings and the answer’s graph embeddings. A system incorporating this novel method is also described in our work. Through comprehensive experimentation using various English and multilingual datasets and two KGs — Freebase and Wikidata — we illustrate the comparative advantage of the proposed method across diverse KG embeddings and languages. This edge is apparent even against robust baseline systems, including seq2seq QA models, search-based solutions and intricate rule-based pipelines. Interestingly, our research underscores that even advanced...

Source

ICCV

StyleDomain: Efficient and Lightweight Parameterizations of StyleGAN for One-shot and Few-shot Domain Adaptation

Aibek Alanov, Vadim Titov, Maksim Nakhodnov, Dmitry P. Vetrov

A*

Source

Domain adaptation of GANs is a problem of fine-tuning GAN models pretrained on a large dataset (e.g. StyleGAN) to a specific domain with few samples (e.g. painting faces, sketches, etc.). While there are many methods that tackle this problem in different ways, there are still many important questions that remain unanswered. In this paper, we provide a systematic and in-depth analysis of the domain adaptation problem of GANs, focusing on the StyleGAN model. We perform a detailed exploration of the most important parts of StyleGAN that are responsible for adapting the generator to a new domain depending on the similarity between the source and target domains. As a result of this study, we propose new efficient and lightweight parameterizations of StyleGAN for domain adaptation. Particularly, we show that there exist directions in StyleSpace (StyleDomain directions) that are sufficient for adapting to similar domains. For dissimilar domains, we propose Affine+ and AffineLight+ parameterizations that allows us to outperform existing baselines in few-shot adaptation while having significantly less training parameters. Finally, we examine StyleDomain directions and discover their many surprising properties that we apply for domain mixing and cross-domain image morphing.

CVPR

Sphere-Guided Training of Neural Implicit Surfaces

Andreea Dogaru, Andrei-Timotei Ardelean, Savva Ignatyev, Egor Zakharov, Evgeny Burnaev

In recent years, neural distance functions trained via volumetric ray marching have been widely adopted for multi-view 3D reconstruction. These methods, however, apply the ray marching procedure for the entire scene volume, leading to reduced sampling efficiency and, as a result, lower reconstruction quality in the areas of high-frequency details. In this work, we address this problem via joint training of the implicit function and our new coarse sphere-based surface reconstruction. We use the coarse representation to efficiently exclude the empty volume of the scene from the volumetric ray marching procedure without additional forward passes of the neural surface network, which leads to an increased fidelity of the reconstructions compared to the base systems. We evaluate our approach by incorporating it into the training procedures of several implicit surface modeling methods and observe uniform improvements across both synthetic and real-world datasets.

A*

Source

Multi-sensor large-scale dataset for multi-view 3D reconstruction

Oleg Voynov, Gleb Bobrovskikh, Pavel Karpyshev, Saveliy Galochkin, Andrei-Timotei Ardelean, Arseniy Bozhenko, Ekaterina Karmanova, Pavel Kopanev, Yaroslav Labutin-Rymsho, Ruslan Rakhimov, Aleksandr Safin, Valerii Serpiva, Alexey Artemov, Evgeny Burnaev, Dzmitry Tsetserukou, Denis Zorin

We present a new multi-sensor dataset for multi-view 3D surface reconstruction. It includes registered RGB and depth data from sensors of different resolutions and modalities: smartphones, Intel RealSense, Microsoft Kinect, industrial cameras, and structured-light scanner. The scenes are selected to emphasize a diverse set of material properties challenging for existing algorithms. We provide around 1.4 million images of 107 different scenes acquired from 100 viewing directions under 14 lighting conditions. We expect our dataset will be useful for evaluation and training of 3D reconstruction algorithms and for related tasks.

A*

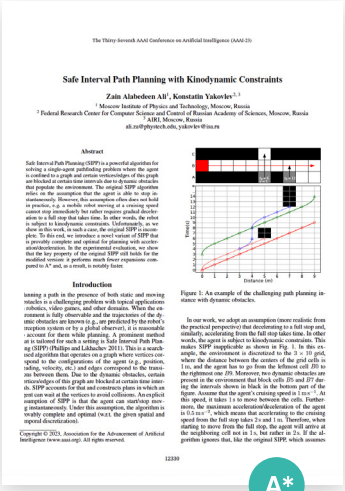
Source

AAAI

Safe Interval Path Planning with Kinodynamic Constraints

Zain Alabedeen Ali, Konstatin Yakovlev

Safe Interval Path Planning (SIPP) is a powerful algorithm for solving a single-agent pathfinding problem where the agent is confined to a graph and certain vertices/edges of this graph are blocked at certain time intervals due to dynamic obstacles that populate the environment. The original SIPP algorithm relies on the assumption that the agent is able to stop instantaneously. However, this assumption often does not hold in practice, e.g. a mobile robot moving at a cruising speed cannot stop immediately but rather requires gradual deceleration to a full stop that takes time. In other words, the robot is subject to kinodynamic constraints. Unfortunately, as we show in this work, in such a case, the original SIPP is incomplete. To this end, we introduce a novel variant of SIPP that is provably complete and optimal for planning with acceleration/deceleration. In the experimental evaluation, we show that the key property of the original SIPP still holds for the modified version: it performs much fewer expansions compared to A* and, as a result, is notably faster.



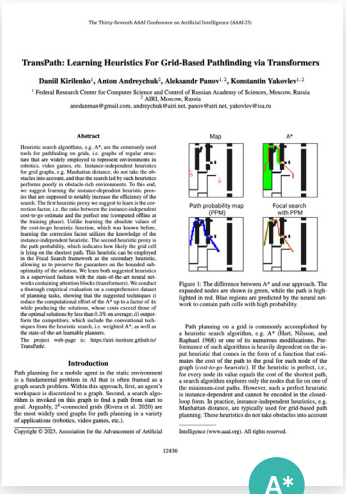
Source

AAAI

TransPath: Learning Heuristics for Grid-Based Pathfinding via Transformers

Daniil Kirilenko, Anton Andreychuk, Aleksandr Panov, Konstantin Yakovlev

Heuristic search algorithms, e.g. A*, are the commonly used tools for pathfinding on grids, i.e. graphs of regular structure that are widely employed to represent environments in robotics, video games, etc. Instance-independent heuristics for grid graphs, e.g. Manhattan distance, do not take the obstacles into account, and thus the search led by such heuristics performs poorly in obstacle-rich environments. To this end, we suggest learning the instance-dependent heuristic proxies that are supposed to notably increase the efficiency of the search. The first heuristic proxy we suggest to learn is the correction factor, i.e. the ratio between the instance-independent cost-to-go estimate and the perfect one (computed offline at the training phase). Unlike learning the absolute values of the cost-to-go heuristic function, which was known before, learning the correction factor utilizes the knowledge of the instance-independent heuristic. The second heuristic proxy is the path probability, which indicates how likely the grid cell is lying on the shortest path. This heuristic can be employed in the Focal Search framework as the secondary heuristic, allowing us to preserve the guarantees on the bounded sub-optimality of the solution. We learn both suggested heuristics in a supervised fashion with the state-of-the-art neural networks containing attention blocks (transformers). We conduct a thorough empirical evaluation on a comprehensive dataset of planning tasks, showing that the suggested techniques i) reduce the computational effort of the A* up to a factor of 4x while producing the solutions, whose costs exceed those of the optimal solutions by less than 0.3% on average; ii) outperform the competitors, which include the conventional techniques from the heuristic search, i.e. weighted A*, as well as the state-of-the-art learnable planners. and, as a result, is notably faster.



Source

AIRI Publications

airi.net/articles/

AIRI Open source repositories

Hugging Face

GitHub

Awards

Aneksandr Panov

A team of researchers from AIRI, MIPT, and FRC CSC RAS, led by Aleksandr Panov and Alexei Staroverov, won first place in an international competition for training robots to navigate indoors.

Alexander Korotin

Alexander Korotin, a researcher at AIRI and a head of the research group at Skoltech, has won an award for pushing the boundaries of knowledge about artificial intelligence.

Ilya Makarov

Three AIRI researchers became winners of the Yandex ML Prize in 2023 — this is the Yandex Prize for scientists and professors in the field of Machine Learning.

Ilya Makarov, head of the “Industrial AI” group, won in the “Scientific Supervisor” category.

Aibek Alanov

Aibek Alanov, a researcher from the “Probabilistic Learning” group, won the “Researchers” nomination.

Anton Razhigaev

FusionBrain group researcher Anton Razhigayev was also honored in the “Researchers” nomination.

Events

AIRI Seminars playlist

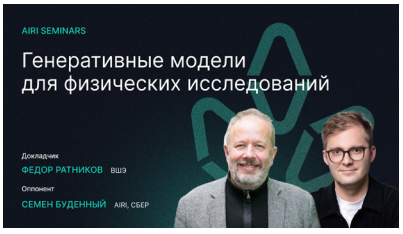
AIRI Seminars

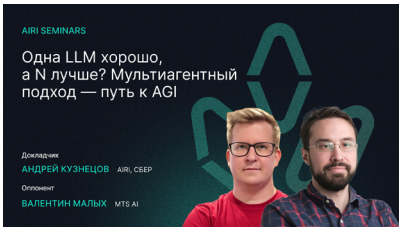
AIRI Seminars is a peer-to-peer scientific dialog that provides the professional community with advances in the field of artificial intelligence.

The seminar is designed to popularize and disseminate the principles and values adhered to by the Institute within AI community, as well as to promote ideas that realize AIRI mission: to create universal AI systems that solve real-world problems.

Leading experts in the field of artificial intelligence from Russia and abroad are invited as speakers or opponents to present and constructively criticize research papers.

This year 21 seminars were held online and offline.





All the scientific seminars are available via AIRI's YouTube channel.

Statistics

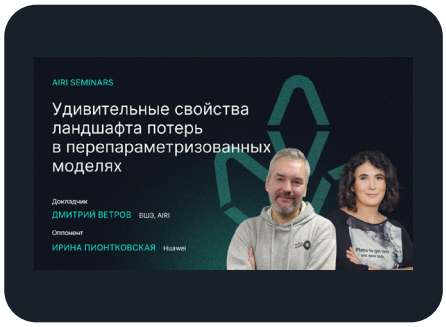
21
seminars

23
organizations

12 758
views

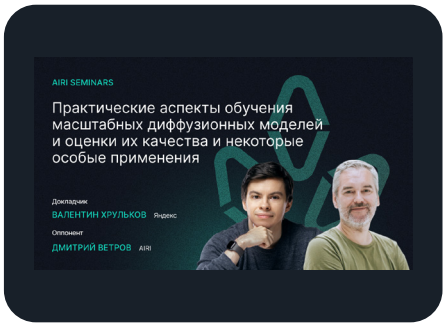
Most viewed seminars

Surprising properties of the loss landscape in overparameterized models



Source

Training of large diffusion models and quality assessment



Source

Team


Aleksandr Panov


Alexey Skrynnik


Alexandra Broytman

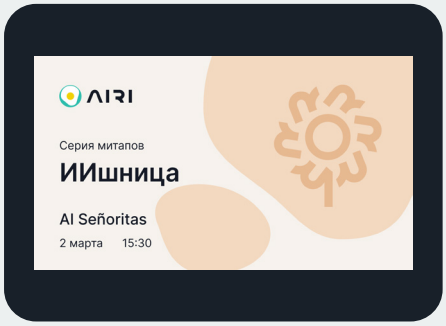

Ekaterina Mamontova


Yuriy Yarovikov


Nikolay Kuschenko

‘ScrAImble’ meetup series

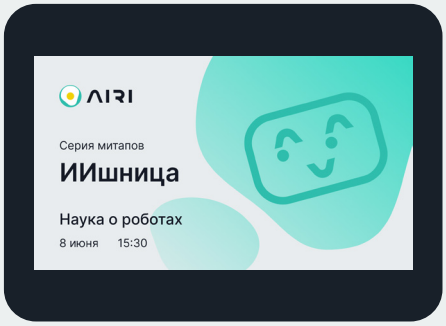
In 2023, we created a special project named ‘ScrAImble’. It consists of a series of online presentations where scientists discuss artificial intelligence in 20-minute scientific reports.



March 2

AI Señoritas

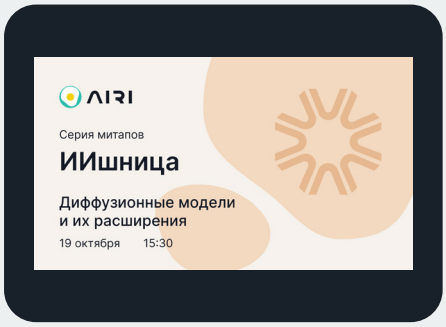
Watch online



June 8

Robotics

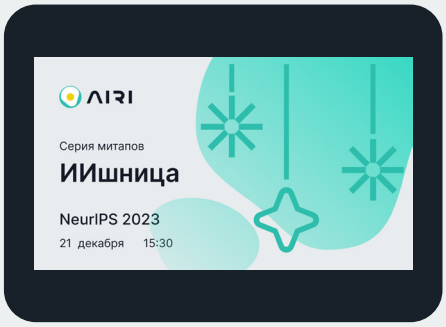
Watch online



October 19

‘Diffusion Models and their Extensions’

Watch online



December 21

NeurIPS 2023

Watch online

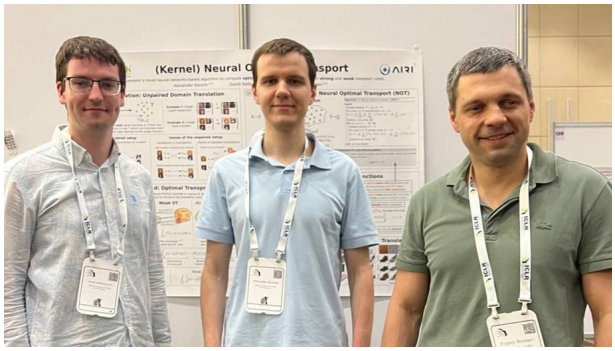
Educational programs

AIRI researchers participated in educational schools, including the HSE Summer School on ML in Bioinformatics, the SMILES Summer School on Machine Learning, and the RAAI 2023 Summer School.

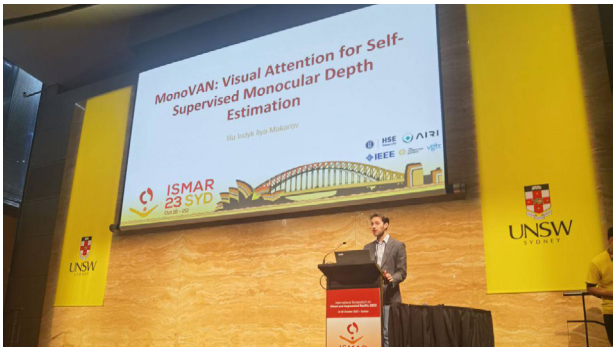
AIRI researchers presented their results at more than 25 scientific conferences

IJCNLP-AACL 2023

EMNLP in Singapore



ICLR conference



IEEE ISMAR

ACL 2023 in Toronto

ICML 2023 in Hawaii

Interspeech 2023 in Dublin

Fall into ML 2023

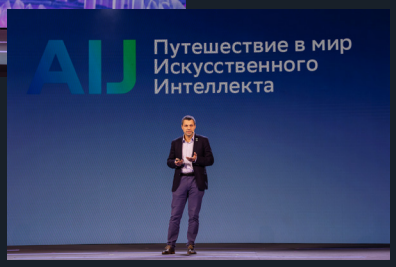
OpenTalks AI 2023 in Yerevan

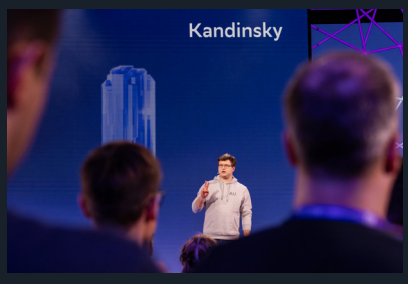
Science 0+

AI Journey 2023

17

researchers gave
presentations
at the conference





AI Journey 2023

14

reserchers presented posters

2

challenges for the AIJ Contest

AIRI researchers prepared two challenges for the AIJ Contest

Strong Intelligence — a task about multimodal chatbots. Rescue AI — a task about detection of chromosomal rearrangements on Hi-C maps.

Science 0+ in China

Ivan Oseledets, Evgeny Burnaev, Semen Budennyy, Konstantin Yakovlev, and Andrey Kuznetsov gave presentations at the AIJ session at the Science 0+ Festival in China.

Young Scientists Congress

On November 29 and 30 at the Young Scientists Congress, AIRI researchers participated in several discussions and presented 3 papers.

Summer with AIRI

Summer with AIRI

The participants were able to apply the knowledge gained during the lectures at practical seminars and in project activities, following which they presented reports on the results of their work.

Participants spent two weeks with leading scientists from AIRI, MIPT, HSE, Skoltech and other reputable research organizations and universities.

700
applications

35
lecturers

80
students

2
weeks

33
projects

132
academic hours

70
scientific posters

Partners

cloud.ru

SBER AI Lab

Skoltech

HSE UNIVERSITY

MIPT

AIRI LEGO club

This year, the informal hobby club of AIRI employees continued to develop actively. Colleagues used to build Lego, get acquainted, invite friends and play quizzes.

Journal club

AIRI researchers hold a journal club to discuss the latest research articles in the field of artificial intelligence.
In 2023, 13 meetings were held.

Partnerships and collaborations

SBER
HEVEL ENERGY GROUP
SOLTEKH
THE GAMALEYA NATIONAL CENTER OF EPIDEMIOLOGY AND MICROBIOLOGY
SBER AI
SBER DEVICES
SBER MED AI
cyberphysics
MIPT
INNOPOLIS UNIVERSITY
HSE UNIVERSITY
Skoltech
AI Research Centre
Sirius
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ РЕВМАТОЛОГИИ ИМЕНИ В.А.НАСОНОВОЙ
FEDERAL RESEARCH CENTER COMPUTER SCIENCE AND CONTROL OF THE RUSSIAN ACADEMY OF SCIENCES
NEIMARK
Russian Academy of Sciences Fersman Mineralogical Museum 1716
MISIS UNIVERSITY
BAUMAN MOSCOW STATE TECHNICAL UNIVERSITY
DIGITAL TECHNOLOGIES IN INDUSTRY ASSOCIATION
ITMO
ITC
АДЫГЕЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ADYGHE STATE UNIVERSITY


AIRI

Contact information

Website

airi.net

Social networks

 airi_research_institute

 artificial-intelligence-research-institute

 AIRIInstitute

 AIRI_inst

 Airi_institute

Address

Moscow, Presnenskaya Embankment 6 build. 2