Aleksandr Drozd

curriculum vitae

(updated August 20, 2024)

Tokyo, Japan
alex@blackbird.pw
blackbird.pw

Brief

Aleksandr Drozd (PhD) works at the intersection of Artificial Intelligence and High Performance Computing, encompassing both academic research and practical applications: using AI in different domains of life and science and developing more performant and scalable AI systems. In addition to that, Aleksandr has an extensive experience in commercial software development and IT consulting.

Employment History

2019.07-onwards	**Research Scientist** at RIKEN Center for Computational Science (R-CCS), High Performance Artificial Intelligence Systems Research Team.
2019.07-onwards	**Visiting Researcher** at Tokyo Institute of Technology, School of Computing
2019.02-2019.06	**Visiting Scientist** at RIKEN Center for Computational Science (R-CCS), High Performance Big Data Research Team.
2018.06-2019.06	**Invited Researcher** at AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL).
2018.04-2019.06	**Researcher** at Tokyo Institute of Technology, School of Computing, Department of Mathematical and Computing Science.
2014.04-2018.03	**Researcher** at Tokyo Institute of Technology, Global Scientific Information and Computing Center.
2005.06-2010.05	**Lecturer / Senior Lecturer** (from 2008) at Moscow State University (Sevastopol Branch, http://sev.msu.ru/), Department of Computational Mathematics and Cybernetics.
2006.09-2009.06	**Software Architect and Developer** at Outsourcing Ukraine (http://www.outsourcing-ukraine.com/).
2005.05-2006.09	**Software Developer** at Soft-Pilot 2000.

Language Proficiency

Japanese: fluent English: fluent Russian: fluent Ukrainian: fluent

Education

2010-2014 Ph.D., Tokyo Institute of Technology, Graduate School of Information Science and Technology. (obtained 2014.03.26)

Thesis title: Memory-Conscious Optimizations for Sorting and Sequence Alignment for Massively Parallel Heterogeneous Architectures.

2000-2005 Specialist degree (M.Sc. equivalent), Moscow State University. Department of Computational Mathematics and Cybernetics.

Thesis title: Semantic Pseudo-Code: Approach to Meaning-Base Search.

Technical Skills

- **Artificial Intelligence**: Development deep learning models for specific tasks: natural language and scientific data processing, machine vision, etc. Working with large language models: training, fine-tuning, integration with applications including retrieval -augmented generation etc. Scaling model training on large distributed systems (up to hundreds of thousands nodes) using hybrid parallelization strategies. Classical machine learning methods. To work with AI models I mainly use PyTorch framework, PyTorch Lightning, specialized packages for language models (Megatron DeepSpeed, langchain, etc), as well as my own solutions.
- **Software Design and Development**: Designing complex architectures, team management, software development processes and DevOps practices like continuous integration and delivery, version control, containers etc.
- **Python:** I started using Python in about 2006 and to this day it is my main tool for prototyping my research ideas, machine learning middle-ware, and back-end development for my web projects and projects of intermediate scale and complexity.
- **High Performance Computing** Parallel programming, using GPUs and other flavours of accelerators. For such tasks I use **C and C++** programming languages, along with specialised frameworks and libraries like CUDA, OpenMP, MPI, OpenCL, TBB etc.
- **Web Development**: I like doing front-end development for my projects myself, preferring minimalist JavaScript frameworks such as HTMX and Alpine JS. I recently use Tailwind CSS and Pico.css for styling, previously having experience with Bulma and Bootstrap. I opt to static site generation when possible, recently using golang-base Hugo SSG, previous python-based Nikola.
- **Cloud Technologies**: working with back-end-as-a-service provides such as Google Firebase, integrating with various APIs (Calendar etc), using infrastracture-as-a-service providers such as AWS, administering self-hosted servers etc.
- Misc: Databases (SQL and noSQL), exploring new programming languages (Go, Kotlin, Zig), using computer algebra and publishing systems etc. In the past I worked a lot with C#. and .net ecosystem.

Recent Projects

AI for optimizing program codes. The most recent project for which I have secured competitive finding from the JSPS KAKENHI program (grant number JP22H03600) is titled "Automated, Scalable, and Machine Learning-Driven Approach for Generating and Optimizing Scientific Application Codes". Computer programs often require labor-intensive manual optimization (re-organization of loops with tiling, fusing, unrolling etc) to perform efficiently on a given hardware target. In this project we develop machine-driven approach to automate software optimization process, particularly focusing of transformation of loops. Focusing on exploring a space of formalized transformations (we are using polyhedral representation to map iteration spaces of loops into multi-dimensional integer sets) allows us to guarantee correctness of generated codes compared to end-to-end generating with language models, popularized recently with models like code-pilot etc.

Developing next national flagship supercomputer. Our center together with industrial partners is currently designing the next machine to succeed supercomputer Fugaku in 2028. My responsibility is to make sure that the new system will excel in running AI workloads: I am coordinating efforts in benchmarking deep learning workloads, surveying emergent trends in development of AI models and methods and evaluating different hardware options, including simulation of not-yet-existing hardware designs.
At the early stages of supercomputer Fugaku development, I had **led an effort for creating deep learning ecosystem** for it. In-depths study of architectures of deep learning frameworks, libraries of specialized performance primitives and intermediate representations allowed us to identified the most promising directions and create large-scale partnership involving Fujitsu ltd, ARM ltd, Linaro ltd. At the later stage of the project, engineering effort was largely moved to Fujitsu ltd.
During my stay in AIST I was involved in a procurement of 4000-GPU "ABCI" supercomputer purpose-built by AIST for AI workloads.
Natural Language Processing: In scientific work on of my focus point is on natural language processing and computational linguistics: from symbolic methods to embeddings to LLMs. During times of popularity of static word embeddings, I led the development of a library for their training and quality assessment (https://vecto.space/). More recently, I have been working on the challenges of interpretability of language models, improving generalization, etc.

I am a **founder and CEO of a company Amigawa GK**, created to help commercialize some of research ideas developed by me and collaborators.
Finally, I stay passionate about **using and further developing my software development skills**. I introduce best software development practices to research work done by me and collaborators, contribute to open-source project and supervise engineering efforts in some of the research projects within the team.
As a **software developer / architect** I was involved in projects like porting software for managing industrial machinery to a different type of real-time operating system; developing an online trading platform for jewelry and gems; small fun projects like generative art library in Python (`https://pycontextfree.blackbird.pw/`) etc.

Selected Fellowships and Grants

- JSPS KAKENHI Grant number JP22H03600 adopted FY 2022: ”Automated, Scalable, and Machine Learning-Driven Approach for Generating and Optimizing Scientific Application Codes”.
- HPCI Project hp210265 “Training Novel Types of Large-Scale Language Models: Tuning”. 1000000 node-hours on supercomputer Fugaku.
- 2021 ABCI Grand Challenge #3: 1000 A-100 GPU/days for scalable weakly supervised video representation learning study.
- HPCI Project hp200281 “Training Novel Types of Large-Scale Language Models: Preparation”. 100000 node-hours on supercomputer Fugaku.
- JSPS KAKENHI Grant number JP17K12739 adopted FY 2017: ”Corpora on Demand: Scalable Methods of Obtaining Linguistic Data”.
- Japanese Government (Monbukagakusho) scholarship for conducting PhD research 2010-2014.

Teaching and Supervision

As a research scientist at the RIKEN Center for Computational Science I have been supervising a number of interns admitted to the RIKEN summer internship program, as well as interns invited individually and supported by ours or external researchg funds. Most recently (September to November 2022) I have been supervising (now Dr.) Giovanny Puccetty from the Scola Superiore Noramle, Pisa. His itnernship resulted in a publication in a prestigious EMNLP conference.

As a post-doctoral appointee at the Tokyo Institute of Technology I have helped advising PhD work of several students (the main supervisor was prof. Satoshi Matsuoka), the most recent is Shweta Salaria, thesis title ”Cross Architecture Performance Prediction”. Courses taught as a lecturer / senior lecturer at the Moscow State University, Faculty of Computational Mathematics and Cybernetics through years 2005-2010:

- **Operating Systems**: Architecture of Unix-like operating systems, inter-process communication mechanisms, C programming language.
- **Object-Oriented Software Design**: C++ programming language, object-oriented approach to software development.
- **Computer Graphics**: basic 2D drawing, 3D projections and transformations, shading, ray tracing.
- **Parallel Data Processing**: Theoretical foundations of parallel computing, OpenMP and MPI libraries/run-times, GPU computing.

I was responsible for developing curricula and teaching materials for these courses, as well as conducting the final examinations. I have also taught fundamentals of computer science at the Faculty of Philology of the Moscow State University.

Community Service

Organizing Academic Events

- **2024 International Workshop on Large Language Models (LLMs) and HPC** in conjunction with CLUSTER 2024 conference. `https://llmhpc.github.io/2024/`
- **The Workshop on Insights from Negative Results in NLP** series (2022, 2023, 2004) In conjunction with *ACL conferences. `https://insights-workshop.github.io`
- **Benchmarking in the Data Center: Expanding to the Cloud** in conjunction with PPoPP 2022: Principles and Practice of Parallel Programming 2022. `https://parallel.computer`
- **International Workshop COmputing using EmeRging EXotic AI-Inspired Systems (CORtEX'22)**. Co-hosted with IPDPS 2022 conference. `https://cortex.ws`
- **"Deep Learning from HPC Perspectives: Opportunities and Challenges"** Mini-Symposium at SIAM PP 2018 conference. `http://meetings.siam.org/sess/dsp_programsess.cfm?SESSIONCODE=63584`
- **"The Third Workshop on Evaluating Vector Space Representations for NLP"**. Co-located with NAACL 2019 conference on June 6 or 7, 2019. `https://repeval2019.github.io`

Chairing/ Program Committee Participation

- I have served as an area chair for the series of EMNLP conferences (efficient NLP track) and program committee member of a number of conferences and workshops, including NAACL, *SEM, SC, ISC, PARCO among others.

Tutorials

- **"Text Representation Learning and Compositional Semantics"**. Tutorial 5 at the 11th Asian Conference on Machine Learning (ACML 2019) `http://www.acml-conf.org/2019/tutorials/drozd-rogers/`
- **"Distributional Compositional Semantics in the Age of Word Embeddings: Tasks, Resources and Methodology"**. Tutorial 4 at LREC 2018 conference. `http://text-machine.cs.uml.edu/lrec2018_t4`

Misc

- From 2017 I serve as one of the organizers of "Tokyo Machine Learning Gym" meetup.
- I actively participate in local meetups and talkathon, recently giving talks at events such as CrossRoads and Tokyo AI Talks etc.

Publications

(1222 citations, h-index=15 as of August 2024)

- Jens Domke, Emil Vatai, Balazs Gerofi, Yuetsu Kodama, Mohamed Wahib, Artur Podobas, Sparsh Mittal, Miquel Pericas, Lingqi Zhang, Peng Chen, Aleksandr Drozd, Satoshi Matsuoka. **At the locus of performance: Quantifying the effects of copious 3D-stacked cache on HPC workloads.** ACM Transactions on Architecture and Code Optimization 20, 2023.
- Satoshi Matsuoka, Jens Domke, Mohamed Wahib, Aleksandr Drozd, Torsten Hoefler. **Myths and legends in high-performance computing.** The International Journal of High Performance Computing Applications 37, 2023.
- Satoshi Matsuoka, Jens Domke, Mohamed Wahib, Aleksandr Drozd, Andrew A Chien, Raymond Bair, Jeffrey S Vetter, John Shalf. **Preparing for the Future—Rethinking Proxy Applications** Computing in Science & Engineering N24 (2), 2022, pp 85-90
- Truong Thao Nguyen, François Trahay, Jens Domke, Aleksandr Drozd, Emil Vatai, Jianwei Liao, Mohamed Wahib, Balazs Gerofi **Why globally re-shuffle? Revisiting data shuffling in large scale deep learning.** 2022 IEEE International Parallel and Distributed Processing Symposium (IPDPS) pp 1085-1096.
- Giovanni Puccetti, Anna Rogers, Aleksandr Drozd and Felice Dell'Orletta. **Outlier Dimensions that Disrupt Transformers are Driven by Frequency** Findings of the Association for Computational Linguistics: EMNLP 2022, pp 1286–1304.
- Prajjwal Bhargava, Aleksandr Drozd, Anna Rogers. **Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics**. Proceedings of the Second Workshop on Insights from Negative Results in NLP (Insights 2021), pp 125–135.
- Steven Farrell, Murali Emani, Jacob Balma, Lukas Drescher, Aleksandr Drozd et al. **MLPerf™ HPC: A Holistic Benchmark Suite for Scientific Machine Learning on HPC Systems.** 2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)
- Jens Domke, Emil Vatai, Aleksandr Drozd, et al. **Matrix Engines for High Performance Computing: A Paragon of Performance or Grasping at Straws?** IPDPS 2021: International Parallel and Distributed Processing Symposium. pp 1056-1065.
- Mohamed Wahib, Haoyu Zhang, Truong Thao Nguyen, Aleksandr Drozd, Jens Domke, Lingqi Zhang, Ryousei Takano, Satoshi Matsuoka. **Scaling distributed deep learning workloads beyond the memory capacity with KARMA**. Proceedings of SC 20: the International Conference for High Performance Computing, Networking, Storage and Analysis. Article No.: 19. pp 1–15.
- Shweta Salaria, Aleksandr Drozd, Artur Podobas, Satoshi Matsuoka. **Learning Neural Representations for Predicting GPU Performance**. International Conference on High Performance Computing 2019, pp 40–58.
- Marzena Karpinska, Bofang Li, Anna Rogers and Aleksandr Drozd. **Subcharacter Information in Japanese Embeddings: When Is It Worth It?** In Proceedings of the Workshop on Relevance of Linguistic Structure in Neural Architectures for NLP (RELNLP) 2018 at ACL 2018. Melbourne, Australia. pp 28–37.

- Bofang Li and Aleksandr Drozd. **Subword-Level Composition Functions for Learning Word Embeddings**. *Proceedings of The 2nd Workshop on Subword and Character level models in NLP (SCLeM)* at NAACL 2018. pp 38–48.
- Shweta Salaria, Aleksandr Drozd, Artur Podobas, Satoshi Matsuoka. **Predicting performance using collaborative filtering** 2018 IEEE International Conference on Cluster Computing (CLUSTER), pp 504–514
- Anna Rogers, Aleksandr Drozd and Bofang Li. **The (too Many) Problems of Analogical Reasoning with Word Vectors**. In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017), Association for Computational Linguistics, pp 135–148, Vancouver, Canada.
- Aleksandr Drozd, Anna Gladkova, Satoshi Matsuoka. **Word Embeddings, Analogies, and Machine Learning: Beyond King - Man + Woman = Queen.** Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp 3519–3530, Osaka, Japan, December 11-17 2016
- Mateusz Bysiek, Aleksandr Drozd and Satoshi Matsuoka. **Migrating Legacy Fortran to Python While Retaining Fortran-Level Performance through Transpilation and Type Hints.** Proceedings of PyHPC 16: the 6th Workshop on Python for High-Performance and Scientific Computing. pp 9-18.
- Anna Gladkova and Aleksandr Drozd. **Intrinsic Evaluations of Word Embeddngs: What Can We Do Better?** in Proceedings of The 1st Workshop on Evaluating Vector Space Representations for NLP, Berlin, Germany, 2016, pp. 36–42.
- Annd Gladkova, Aleksandr Drozd and Satoshi Matsuoka. **Analogy-based Detection of Morphological and Semantic Relations With Word Embeddings: What Works and What Doesn't.** Proceedings of NAACL-HLT-SRW 2016, pp 8–15.
- Aleksandr Drozd, Anna Gladkova, Satoshi Matsuoka. **Discovering Aspectual Classes of Russian Verbs in Untagged Large Corpora.**The 2015 IEEE International Conference on Data Science and Data Intensive Systems (DSDIS 2015), At Sydney, Australia, Dec 2015, pp 61 - 68.
- Aleksandr Drozd, Anna Gladkova, Satoshi Matsuoka. **Python, Performance and Natural Language Processing**. 5th Workshop on Python for High-Performance and Scientific Computing, at Austin, Texas, USA, Nov 2015 in conjunction with SC15, pp 1-10.
- Aleksandr Drozd, Olaf Witkowski, Satoshi Matsuoka, Takashi Ikegami.**Signal-Driven Swarming: A Parallel Implementation of Evolved Autonomous Agents to Perform A Foraging Task.** Proceedings of SWARM 2015 - The First International Symposium on Swarm Behavior and Bio-Inspired Robotics, Kyoto, Oct 2015.
- Aleksandr Drozd, Naoya Maruyama and Satoshi Matsuoka. **Sequence Alignment on Massively Parallel Heterogeneous Systems**, IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum. 2012, Shanghai, China. Proceedings of IPDPS 12 workshops, pages 2498 - 2501
- Aleksandr Drozd and Satoshi Matsuoka. **A Multi GPU Read Alignment Algorithm with Model-based Performance Optimization**, 10th International Conference, on High Performance Computing for Computational Science - VECPAR 2012, Kobe, Japan,

July 17-20, printed asSpringer's Lecture Notes in Computer Science N7851, pages 270-277.

- Bofang Li, Aleksandr Drozd, Yuhe Guo, Tao Liu, Satoshi Matsuoka, Xiaoyong Du. **Scaling Word2Vec on Big Corpus** Data Science and Engineering, June 2019, Volume 4, Issue 2, pp 157–175.
- Hideyuki Shamoto, Koichi Shirahata, Aleksandr Drozd, Hitoshi Sato and Satoshi Matsuoka. **GPU-Accelerated Large-Scale Distributed Sorting Coping with Device Memory Capacity**. IEEE Trans. Big Data 2(1): 57-69 (2016)
- Aleksandr Drozd, Olaf Witkowski, Satoshi Matsuoka and Takashi Ikegami. **Critical Mass in the Emergence of Collective Intelligence: a Parallelized Simulation of Swarms in Noisy Environments**. Artificial Life and Robotics 2016, volume 21, number 3, pp 317-323
- Anna Gladkova and Aleksandr Drozd, **Towards Easier Querying of XML-based Linguistic Corpora,** Taurida Bulletin of Mathematics and Informatics. #2, 2009, pages 71-77 http://tvim.info/node/146