Muhammed Fatih Balin

APPLIED SCIENTIST II · PH.D. CANDIDATE · HPC AND ML PRACTITIONER

Santa Clara, CA 95054, USA

☎ (+1) 470-645-7922 | ✉ m.f.balin@gmail.com | mfbal.in | mfbalin | mfbalin

“Strive always to excel in virtue and truth.”

Areas of Interest

- Large scale deep learning on graphs via graph neural networks and its variants
- Parallel algorithms for CPUs and GPUs, distributed algorithms for clusters of computers
- Applying deep learning to discrete optimization problems

Education

Georgia Institute of Technology

Atlanta, GA, USA

PH.D IN COMPUTATIONAL SCIENCE AND ENGINEERING — *Aug. 2019 - PRESENT*

- Focusing on High Performance Computing and Machine Learning advised by Umit V. Catalyurek.

Bogazici University

Istanbul, Turkey

B.S. IN COMPUTER ENGINEERING AND MATHEMATICS, (DOUBLE MAJOR), GPA: 3.94/4.00 — *Sep. 2015 - Jun 2019*

- Rank of 1 in department, 3 in the college among 400 engineering students.

Columbia University in the City of New York

New York, NY, USA

NON-DEGREE EXCHANGE STUDENT IN COMPUTER SCIENCE, GPA: 4.07/4.33 — *Jan. 2018 - May. 2018*

- Rank of 4 among 250 students in a graduate level theory-heavy machine learning course.

Honors & Awards

2019	**Summa Cum Laude**, Bogazici University	*Istanbul, Turkey*
2018	**Dean’s List**, Columbia University in the City of New York	*New York, NY, USA*
2017	**57th place**, Google Hashcode	*Online*
2017	**7th place**, ACM-ICPC SEERC Coding Competition	*Vinnytsia, Ukraine*
2016	**2nd place**, Istanbul Technical University - IEEE Coding Competition	*Istanbul, Turkey*

Publications

- Kaan Sancak, Zhigang Hua, Jin Fang, Yan Xie, Bo Long, Andrey Malevich, **M. F. Balin**, U. V. Catalyurek, ”A Fast and Effective Alternative to Graph Transformers”, AAAI Conference on Artificial Intelligence, Mar 2025.
- **M. F. Balin**, Dominique LaSalle, U. V. Catalyurek, ”Cooperative Minibatching in Graph Neural Networks”, Transactions on Machine Learning Research (TMLR), Jan 2025.
- Kaan Sancak, **M. F. Balin**, U. V. Catalyurek, ”Do We Really Need Complicated Graph Learning Models? – A Simple but Effective Baseline”, Learning on Graphs Conference (LoG), Nov 2024.
- Vaibhav Sharma, Abhinav Nagpal, **M. F. Balin**, ”SIRD: Symbolic Integration Rules Dataset”, 3rd MATH-AI Workshop at NeurIPS, Dec 2023.
- **M. F. Balin**, U. V. Catalyurek, ”Layer-Neighbor Sampling – Defusing Neighborhood Explosion in GNNs”, Neural Information Processing Systems (NeurIPS), Dec 2023.
- **M. F. Balin**, X. An, A. Yasar, U. V. Catalyurek, ”A Novel Subgradient-based Method for d-Dimensional Rectilinear Partitioning”, Technical Report, Oct 2023.
- **M. F. Balin**, K. Sancak, U. V. Catalyurek, ”MG-GCN: Scalable Multi-GPU GCN Training Framework”, International Conference on Parallel Processing (ICPP), Aug 2022.
- A. Yasar, **M. F. Balin**, X. An, K. Sancak, U. V. Catalyurek, ”On Symmetric Rectilinear Matrix Partitioning”, Journal of Experimental Algorithmics (JEA), Sep 2022.
- M. Y. Ozkaya, **M. F. Balin**, A. Pinar, U. V. Catalyurek, ”A scalable graph generation algorithm to sample over a given shell distribution”, IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Workshop on Graphs, Architectures, Programming, and Learning, May 2020.
- **M. F. Balin**, A. Abid, J. Zou, ”Concrete Autoencoders for Differentiable Feature Selection and Reconstruction”, International Conference on Machine Learning (ICML), June 2019.

Talks

CSE 6230: HPC Tools and Applications — *Atlanta, GA, USA*

GUEST LECTURER FOR GRAPHBOLT — *Mar. 2024*

- Taught about the open-source GraphBolt library as a lecture on scalable GNN training in the HPC Tools and Applications class.

37th Conference on Neural Information Processing Systems (NeurIPS) — *New Orleans, LA, USA*

PRESENTER FOR LAYER-NEIGHBOR SAMPLING — DEFUSING NEIGHBORHOOD EXPLOSION IN GNNS — *Dec. 2023*

- Presented our paper via an online talk on scalable GNN training in the NeurIPS conference.

The Platform for Advanced Scientific Computing (PASC) Conference — *Basel, Switzerland*

PRESENTER FOR MG-GCN: SCALABLE MULTI-GPU GCN TRAINING FRAMEWORK — *Jun. 2022*

- Presented our paper on scalable GNN training in the High-Performance ML: Scale and Performance Minisymposium

International Conference on Machine Learning 2019 — *Los Angeles, CA, USA*

PRESENTER FOR CONCRETE AUTOENCODERS FOR DIFFERENTIABLE FEATURE SELECTION AND RECONSTRUCTION — *Jun. 2019*

- Presented our paper on unsupervised feature selection in the unsupervised learning session

Research Experience

Fatima Fellowship — *Atlanta, GA, USA*

MENTOR — *Apr. 2021 - PRESENT*

- Designing novel algorithms for GNN training with two mentees from underrepresented parts of the world.
- Tackling symbolic integration of functions powered by A.I. to speed up the search process with two other mentees.

TDAlab, Georgia Institute of Technology — *Atlanta, GA, USA*

GRADUATE RESEARCH ASSISTANT — *Aug. 2019 - Aug 2024*

- Working on speeding up Stochastic Gradient Descent on GNNs, by coming up with new sampling algorithms and taking advantage of the GNN batch dynamics. Currently multiple manuscripts are under review.
- Scaled Graph Convolutional Network Training to multi-GPU, accepted by ICPP 2022
- Proposed subgradient optimization for rectilinear partitioning of sparse matrices and point datasets
- Devised and implemented a shared memory code and a hybrid version with MPI using C++11 threads, lock-free data structures and fine-grained parallelism to generate random graphs given a k-core structure, accepted at IPDPSW 2020
- Implemented a 2D sparse prefix sum data structure using persistent Binary Indexed Trees to speed up the prefix sum queries in matrix partitioning algorithms, proposed sparsification idea to make queries much faster and proposed more efficient heuristics for the symmetric rectilinear partitioning problem, accepted at JEA 2021

Bogazici University & Stanford University — *Istanbul, Turkey & Stanford, CA, USA*

UNDERGRADUATE REMOTE COLLABORATOR — *Sep. 2018 - Mar. 2019*

- Introduced a new scalable unsupervised feature selection algorithm called Concrete Autoencoders based on the Concrete Distribution and Autoencoders.
- Published and presented our work at ICML 2019 as a first author, paper is available at arXiv:1901.09346, talk is available on slideslive.

Creative Machines Lab, Columbia University — *New York, NY, USA*

UNDERGRADUATE RESEARCH ASSISTANT — *Jan. 2018 - May. 2018*

- Devised and implemented a topology optimization algorithm in C++ via simulated annealing and a simulation approach.
- Implemented a visualization tool in C++ and OpenGL for the optimization and simulation algorithm.

Teaching Experience

Georgia Institute of Technology — *Atlanta, GA, USA*

TEACHING ASSISTANT FOR CSE6740 - COMPUTATIONAL DATA ANALYTICS — *Aug. 2020 - Dec. 2020*

- Held office hours, graded assignments and prepared exams.

Georgia Institute of Technology — *Atlanta, GA, USA*

TEACHING ASSISTANTSHIPS FOR CSE6010 - COMPUTATIONAL PROBLEM SOLVING — *Aug. 2019 - Dec. 2019*

- Gave 3 of the lectures, held office hours and graded programming assignments.

TUBITAK — *Afyon, Turkey*

ADVISOR & INSTRUCTOR FOR OLYMPIADS IN INFORMATICS — *Sep. 2017 & Sep 2018*

- Gave lectures on discrete mathematics, advanced data structures and algorithms.
- Mentored high school students in better preparing for the Olympiads in Informatics.

Work Experience

Amazon Web Services
Santa Clara, CA

Applied Scientist II
Oct. 2024 - PRESENT

- Adding quantized data type support for Amazon's in-house AI accelerators.
- Inventing quantization algorithms for fast AI inference.
- Guiding the architecture team on future generation quantization support.

Deep Graph Library (DGL)
Remote

Individual Contributor and Technical Lead
Aug. 2022 - PRESENT

- Leading the design and implementation of the new multi-GPU GNN dataloading library GraphBolt.
- Contributed CPU and optimized CUDA implementation of Layer-Neighbor Sampling algorithm for Graph Neural Networks.
- Exploring Cooperative Minibatching for GNNs for multi-GPU systems.
- Contributed GPU Embedding cache into the core library.
- Miscellaneous improvements and optimizations in general.

NVIDIA
Santa Clara, CA

Developer Technology Intern - AI
May. 2023 - Aug. 2023

- Implemented fused fine-grained FP8 quantization kernels in TensorRT-LLM for Hopper GPUs. It is almost as fast as static quantization while requiring no calibration, enabling quantizing any LLM model on the fly.

NVIDIA
Santa Clara, CA

Developer Technology Intern - AI
May. 2022 - Aug. 2022

- Worked on developing a system based on DGL for Cooperative GNN training. Also developed a vertex cache embedding system to speedup embedding transfers from CPU to GPU.
- Continued to develop a new GNN sampling algorithm called LABOR, aiming to contribute to the DGL framework.

Pacific Northwest National Laboratory
Richland, WA

Research Intern
May. 2021 - Aug. 2021

- Developed distributed data structures and algorithms on the open source distributed programming framework SHAD.

Icron Technologies
Istanbul, Turkey

Research Engineering Intern
Jul. 2017 - Aug. 2017

- Learned about applications of optimization techniques such as mixed integer programming in industry.
- Learned Icron, a fully functional visual programming language developed and being used at Icron Technologies.

Baykar Technologies
Istanbul, Turkey

Software Engineering Intern
Jul. 2016 - Sep. 2016

- Implemented a library using suffix arrays to handle search queries efficiently for a GUI application.
- Implemented a tool to convert simple C header and source files to C#.
- Implemented a line of sight algorithm to determine where a Unmanned Aerial Vehicle can go without losing line of contact with a receiver.

Baykar Technologies
Istanbul, Turkey

Software Engineering Intern
Jun. 2015 - Aug. 2015

- Worked on a GUI application that monitors and controls an Unmanned Aerial Vehicle's state to add new features and fix bugs.
- Found bottlenecks in code processing post-flight data to get a 50x of speedup.

Skills

- **Spoken Languages:** English (TOEFL 111), German (Abitur), Turkish
- **Technology:** C++23, MPI, Python, Tensorflow, Pytorch, DGL, OpenCV, Java, Matlab, SyCL, CUDA, Parallel Programming
- **Certificates:** Entrepreneurship Seminar Series - BIC Angels, Istanbul, Turkey, 2016