Maksim Levental

maksim.levental@gmail | makslevental.github.io | github.com/makslevental

INTERESTS

Broadly, I'm interested in using, improving and designing application-specific hardware accelerators for Deep Neural Networks (DNNs), for science applications. Specifically, using compiler techniques such as symbolic execution, and mathematical optimization techniques such as integer programming, to improve memory consumption and inference latencies of DNNs on GPUs, FPGAs, and ASICs.

EDUCATION

Ph.D., **Computer Science**, *University of Chicago*, GPA: 4.0	**Oct. 2020 — Present**
M.S., **Computer Science**, *University of Florida*, GPA: 3.7	**Aug. 2014 — Dec. 2015**
B.S., **Pure Math**, **Honors**, *Florida State University*, GPA: 3.7	**Aug. 2007 — Dec. 2010**

EXPERIENCE

Graduate Research Assistant, Globus Labs, *University of Chicago* **Oct. 2020 — Present**

- Investigated a design workflow that allows exploring algorithmically complex hardware designs and developing reusable hardware libraries for the needs of scientific instruments at the edge. Specifically, investigated Chisel and other HDLs for designing DNNs and compression algorithms.
- Helped develop a workflow that connects the Data and Learning Hub for Science (DLHub), a repository for publishing AI models, with the Hardware-Accelerated Learning (HAL) cluster, using funcX as a universal distributed computing service. Specifically, deployed Autoencoder models for identifying binary black hole mergers in LIGO data.
- Developed a hybrid CPU-GPU Difference of Gaussians method for characterizing the volume distribution of unburnt nanoparticle synthesis solution, to enable near-real-time optimization and steering of Flame Spray Pyrolysis experiments.

Research Intern, PyTorch, *Facebook* **June 2021 — Present**

- Developed functionality for statically allocating memory for reduced inference latency in production and OSS PyTorch models.
- Developed symbolic analysis techniques for inferring memory requirements of intermediate tensors in dynamic neural networks. Specifically, used and extended shape analysis in TorchScript, combined with MIP formulation of storage allocation, to derive upper bounds on sizes of intermediate tensors.

Teaching Assistant, Computer Science Department, *University of Chicago* **Apr. 2021 — June 2021**

- Assisted with grading, designed assignments, and hosted discussions for MPCS 52040 - Distributed Systems. Specifically, contributed example implementation of RAFT consensus in Rust.

Ranking Engineer Intern, Groups, *Facebook* **June 2020 — Sep. 2020**

- Investigated effects of repeated recommendations on conversion for Groups You Should Join (GYSJ). Specifically, developed and shipped pipelines and metrics for measuring repetition.
- Developed features used in scoring recommendations, including collecting, cleaning, and transforming data.
- Ran A/B tests on 100MM population to measure efficacy of aforementioned features in reducing repetition and increasing conversion. Effects were marginal (i.e., not statistically significant).

Graduate Research Assistant, *University of Florida* **Aug. 2014 — Dec. 2015, Aug. 2018 — Dec. 2019**

- Investigated classical and DNN Super Resolution techniques for improving object detection in Long Wave Infrared imagery. Specifically, improved detection AUC for very low-resolution targets by 5% (using SRGAN).
- Developed a novel training paradigm using, region-based stratified cross-validation, to improve learning induction across disparate (in terms of geography) landmine data sets.
- Researched using Gaussian Processes (GP) for landmine detection. Specifically, modeling soil permittivity/dielectric changes as GPs.

- Implemented Postgres backend for landmine detection algorithm/testing/scoring framework (TUF).

Peace Corps Education Volunteer, *Mbale, Uganda*   **March 2011 — March 2013**
- Taught O level and A level math classes to rural school children. Ran study sessions for national exams.
- Taught IT classes to students and teachers on standard software suites (Office, Creative Suite, etc.). Worked on initiative to bring WiFi to the school campus.
- Taught classes on life skills, science at girls (GLOW) and boys (BUILD) empowerment camps.

SKILLS

Formal Languages	Python, C++, Rust, SQL, CUDA, Scala
Human Languages	English (native), Russian (native)
Platforms	LLVM, MLIR, PyTorch, TensorFlow, PostgreSQL
Skills	Deep Learning, Compilers, Hardware
Coursework	*Computer Science*: Quantum Computing, Deep Learning Systems, Analysis of Algorithms, Consensus and Economics, Automata
	Math: Statistics, Numerical Linear Algebra, Optimization, Real Analysis, Topology, Algebra, PDEs
	Physics: Computational Physics, Electricity and Magnetism, Quantum Mechanics, Statistical Mechanics, Nuclear Physics, Waves and Optics

PUBLICATIONS

Leventa M., Bao B., Chard K., Foster I. *MemoMalloc: Memory Planning for Deep Neural Networks.* Submitted to 16th USENIX Symposium on Operating Systems Design and Implementation (OSDI '22).

Huerta E., Khan A., Huang X., Tian M., **Leventa M.**, Chard R., Wei W., Heflin M., Katz D., Kindratenko V., Mu D., Blaiszik B., Foster I. *Accelerated, Scalable and Reproducible AI-driven Gravitational Wave Detection.* Nature Astronomy volume 5, pages 1062–1068 (2021).

Yoshii K., Sankaran R., Strempfer S., **Leventa M.**, Hammer M., Miceli A. *A Hardware Co-design Workflow for Scientific Instruments at the Edge.* Accepted to the Smoky Mountains Computational Sciences and Engineering Conference (SMC '21).

Leventa M., Chard R., Libera J. A., Chard K., Koripelly A., Elias J., Schwarting M., Blaiszik B., Stan M., Chaudhuri S., Foster I. *Towards Online Steering of Flame Spray Pyrolysis Nanoparticle Synthesis.* **Best Paper at 2020 IEEE/ACM 2nd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP '20).**

Wilson J., Toska F., **Leventa M.**, Dobbins P. *A Deep Neural Network Model for Hazard Classification.* Artificial Intelligence and Machine Learning in Defense Applications (2019).

TECHNICAL REPORTS

Leventa M. *Tensor Networks for Simulating Quantum Circuits on FPGAs.* arXiv:2108.06831 (2021).
Leventa M., Orlova E. *Comparing the Costs of Abstraction for DL Frameworks.* arXiv:2012.07163 (2020).