Samuel S. Ogden

Current position

PhD Candidate (ABD),
The Cake Lab,
Worcester Polytechnic Institute

Cell: 802-558-2650
Email: Samuel.S.Ogden@gmail.com
URL: http://samogden.net/
GitHub: github.com/samogden

Research interests

My research interests focus on enabling deep learning to be used in real-world scenarios in as seamless a way as possible. This translates to my work focusing on designing systems to execute deep learning models on and for resource constrained devices, such as a mobile phones, through machine learning, performance modeling, and cloud computing.

Education

2022 (ABD) PhD in Computer Science, Worcester Polytechnic Institute, Worcester MA
Dissertation: *Mobile-Oriented Deep Learning Inference: Fine- and Coarse-grained Approaches*
Advisor: Dr. Tian Guo
Committee Members:
- Dr. Emmanuel Agu (Worcester Polytechnic Institute),
- Dr. Craig Shue (Worcester Polytechnic Institute),
- Dr. Xiangnan Kong (Worcester Polytechnic Institute),
- Dr. Yue Cheng (George Mason University)

2013 MS in Computer Science, University of Vermont
2010 BS in Electrical Engineering, Pure Mathematics, University of Vermont

Publications

Peer Reviewed

[7] Samuel S. Ogden, Guin R. Gilman, Robert J. Walls, Tian Guo, (2021), "Many Models at the Edge: Scaling Deep Inference via Model-Level Caching" (10 pages), *2nd IEEE International Conference on Autonomic Computing and Self-Organizing Systems* (**ACSOS'21**) (Acceptance Rate 23%)

[6] Samuel S. Ogden, Xiangnan Kong, Tian Guo, (2021), "PieSlicer: Dynamically Improving Response Time for Cloud-based CNN Inference" (8 pages), *12th ACM/SPEC International Conference on Performance Engineering* (**ICPE'21**) (Acceptance Rate 29%)

[5] Guin R. Gilman, Samuel S. Ogden, Tian Guo, Robert J. Walls, (2020), "Demystifying the Placement Policies of the NVIDIA GPU Thread Block Scheduler for Concurrent Kernels" (7 pages), *38th International Symposium on Computer Performance, Modeling, Measurements and Evaluation* (**PERFORMANCE'20**) (Acceptance Rate 23.5%)

[4] Samuel S. Ogden, Tian Guo, (2020), "MDInference: Balancing Inference Accuracy and Latency for Mobile Applications" (11 pages), *IEEE International Conference on Cloud Engineering* (Invited) (**IC2E'20**) (Acceptance rate 51%)

[3] Guin R. Gilman, Samuel S. Ogden, Robert J. Walls, Tian Guo, (2019), "Challenges and Opportunities of DNN Model Execution Caching", (5 pages) *MiddleWare DIDL Workshop* (**DIDL'19**)

[2] Tian Guo, Robert J. Walls, Samuel S. Ogden, (2019), "EdgeServe: Efficient Deep Learning Model Caching at the Edge" (3 pages), *4th ACM/IEEE Symposium on Edge Computing* (**SEC'19**)

[1] Samuel S. Ogden, Tian Guo, (2018), "MODI: Mobile Deep Inference Made Efficient by Edge Computing" (7 pages), *USENIX Annual Technical Conference HotEdge Workshop 2018* (**HotEdge'18**)

Preprints

[3] Samuel S. Ogden, Tian Guo, "Characterizing the Deep Neural Networks Inference Performance of Mobile Applications", *arxiv.org/1909.04783*

[2] Samuel S. Ogden, Tian Guo, "ModiPick: SLA-aware Accuracy Optimization for Mobile Deep Inference"*arxiv.org/1909.02053*

[1] Samuel S. Ogden, Tian Guo, "CloudCoaster: Transient-aware Bursty Datacenter Workload Scheduling"*arxiv.org/1907.02162*

Invited Talks

[1] Samuel S. Ogden, "Mobile Deep Inference", Guest Lecture, *Mobile and Ubiquitous Computing*, Worcester Polytechnic Institute, 2021

Campus Presentations

[3] "CacheRipper: A Content Delivery Network for Deep Learning Models", 2021 (5 min talk), *Annual Graduate Research Innovation Exchange*, Worcester Polytechnic Institute

[2] "MDInference: Balancing Inference Accuracy and Latency for Mobile Applications", 2020 (Poster), *Annual Graduate Research Innovation Exchange*, Worcester Polytechnic Institute

[1] "ModiServe: Efficient SLA-aware Mobile Deep Inference Serving Platform", 2020 (Poster), *Annual Graduate Research Innovation Exchange*, Worcester Polytechnic Institute

Research experience

2017-present **Graduate Research Assistant**, Worcester Polytechnic Institute (PI: Dr. Tian Guo)
Deep Neural Network Execution Caching

- Goal: Approach serving deep learning models from a caching perspective in order to reduce resource requirements and improve resource utilization.
- Core challenges: Number of unique models to host, scale of workload, size of models.
- Results: By considering both model and workload characteristics we show that using CPU-based servers is more cost-effective with minimal impact on latency. Further, by treating models as cacheable objects we can reason about their relative utility and improve resource utilization by adding or removing them from memory as needed. This allows for using

smaller servers with less expensive components, thus allowing us to move model inference servers closer to the edge.
- Project website: https://cake.wpi.edu/ripcord/
- Code: https://github.com/cake-lab/CremeBrulee
- Publications: **ACSOS'21**

Per-request Inference Optimization
- Goal: Consider the characteristics of each inference execution in order to improve its performance, both in terms of latency and accuracy, as much as possible.
- Core challenges: Inherently constrained and variable mobile environment.
- Results: Through modeling of execution and pre-execution steps we demonstrated that it is possible to change execution pipeline decisions at runtime leading to improved, and bounded, latency while opportunistically improving accuracy.
- Code: https://github.com/cake-lab/PieSlicer, https://github.com/cake-lab/MDInference
- Publications: **ICPE'21**, **IC2E'20**

On-devince Inference Exploration
- Goal: Explore the possibility of on-device inference using deep learning models.
- Core challenges: Model size, model resource constraints, model optimization techniques
- Results: We found that different optimizations affected models in a range of ways, with trade-offs being made between model size, execution latency, and accuracy. Overall, leveraging deep learning models on-device is feasible, but due to the overhead and constrained resources remote inference is a viable alternative, especially on older and less capable devices.
- Code: https://github.com/cake-lab/Mobile-deep-inference-benchmark-app
- Publications: **HotEdge'18**

Transient-Aware Workload Scheduling.
- Goal: Reduce cost of workload execution by leveraging transient resources.
- Core challenges: Transient resources are inherently unreliable while workload scheduling relies heavily on resource availability and reliability.
- Results: We showed that by leveraging transient resources, and being aware of their unreliability when scheduling, it was possible to reduce workload execution time while maintaining or lowering cost.

2010-2012 **Graduate Research Assistant**, University of Vermont (PI: Dr. Christian Skalka)
SnowFlake: Wireless sensor network for estimation of snowpack depth and groundwater.

2009-2010 **Senior Capstone Project**, University of Vermont (External Partner: MITRE Corp.)
makeONE: Exploration project focused on capabilities of 3D printers and development a database of replacement parts for military and civilian users.

Teaching experience

Graduate Teaching Assistant, Worcester Polytechnic Institute
Accelerated Object-Oriented Design Concepts (2021),
- Course Description: Freshman level accelerated introduction to OOP targeting students who have previously had some programming experience.
- Size: 101 students, 1 TA, 4 undergraduate student assistants.
- Extra Responsibilities: As the sole TA, I was responsible for running exam review sessions, setting up one-on-one help sessions with struggling students, plurality of grading, and tracking of logistics.

Mobile & Ubiquitous Computing (2021),

- Course Description: Senior-level mobile application design course, introducing key concepts of mobile oriented design as well as modern programming techniques.
- Size: 82 students, 2 TAs
- Extra Responsibilities: I designed a tutorial for the class, as well as giving a guest lecture on the challenges and recent research into integrating deep learning models into mobile applications.

Introduction to Artificial Intelligence (2017),

- Course Description: Senior-level overview of artificial intelligence, from early expert systems to modern deep learning techniques.
- Size: 87 Students, 2 TAs
- Extra Responsibilities: Designed and ran term-long project focused on the development of Gomoku engines that culminated in an extra credit in-class tournament that generated significant interest and engagement.

Computer Networks (2017),

- Course Description: Junior-level introduction to computer networking, including low-level implementations to high-level network communication approaches.
- Size: 50 students, 1 TA
- Extra Responsibilities: Setting up one-on-one help sessions with struggling students

Graduate Teaching Assistant, University of Vermont

Puzzles, Games, and Algorithms (2011),

- Course Description: A freshman-level course aimed at non-majors to introduce them to computer science and the use of algorithms through a series of puzzles games and algorithms including maze solving, code breaking, and unscrambling rubrics cubes.
- Extra Responsibilities: Designed and ran a weekly 3-hour lab aimed to augment in-class lectures by teaching concrete programming techniques to write algorithms.

Introduction to Web Design (2011),

- Course Description: Sophomore-level introduction to web design course.

Teaching Assistant, University of Vermont

Introduction to Engineering,

- Course Description: Freshman-level overview of engineering disciplines and techniques designed to introduce students to the field and help them experience a range of possible careers.
- Extra Responsibilities: Helped to design semester-long project and then advised and assisted students in the design and implementation of their solutions.

Industry experience

2012-2017 Principal Engineer, GLOBALFOUNDRIES/IBM, Essex VT

- Core responsibilities: Design and develop tools for ASIC verification, lead daily team meetings, interface with external clients and internal teams to enforce design and deadline constraints, develop new technology nodes

2009 Lab Technician, University of Vermont, Burlington VT

- Core responsibilities: Setup experimental stations, identify and reorder necessarily supplies, repair student computer stations

Research Mentoring

2020-2021 Justin Aquilante, Ann Jicha, "Heterogeneity Aware Federated Learning", *Computer Science Major Qualifying Project* (Senior Capstone)

- Project description: The goal of this project was to explore the impacts of different training methods for federated deep learning, with particular attention to heterogeneous devices and data with Dr. Tian Guo, would be seen in a mobile environment.
- Key Responsibilities: I co-advised the students on their project, both in terms of direction and implementation. This project covered 21 weeks with biweekly status and brainstorming meeting. Further, I helped them approach challenges that they encountered and together we came up with plans to implement state-of-the-art techniques for testing and comparison.

Grants, honors & awards

2019,2020 Computer Science Graduate Leadership Award
2010-2011 Vermont Space Grant Consortium GRA Recipient
2007-2010 Science and Mathematics Access to Retain Talent Grant
2006-2007 Academic Competitiveness Grant
2008 Putnam Exam Score of 12
2006 AP Scholar with Distinction

Service

2019-present Dean's Graduate Student Advisory Council
2019-present Computer Science Graduate Council
2020 Graduate Student Government Representative
2021-present Climbing Team Graduate Advisor
2018-2021 Outing Club Graduate Advisor
2019-present Student Development and Counseling Center Student Support Network

Membership

2018-present IEEE
2018-present ACM

Skills

Machine Learning, Statistical Methods, Cloud and Edge Computing, Network Security, Mobile Device Oriented Programming, Fundamentals of Scientific Teaching and Pedagogy