Workshop organizers make last-minute changes to their schedule. Download this document again to get the lastest changes, or use the MLSys mobile application.

Schedule Highlights

March 4, 2020

Ballroom A, On-Device Intelligence Chandra, Warden, Venkatesh, Lin

Level 1 Room 3, SARA: Secure and Resilient Autonomy Bose, Chandramoorthy, Vega, Swaminathan

Level 3 Room 10, CANCELED: Automated Machine Learning For Networks and Distributed Systems Arzani, Darvish Rouhani

Level 3 Room 5, MLOps Systems Dutta, Zaharia, Zhang

Level 3 Room 6, Benchmarking Machine Learning Workloads on Emerging Hardware St John, Emani

Level 3 Room 8, Resource-Constrained Machine Learning (ReCoML 2020) Ben Itzhak, Narodytska, Aberger

Level 3 Room 9, Software-Hardware Codesign for Machine Learning Workloads Gupta, Wohlbier, Low, Vetter, Vassilieva
On-Device Intelligence

Vikas Chandra, Pete Warden, Ganesh Venkatesh, Yingyan Lin

Ballroom A, Wed Mar 04, 09:00 AM

AI has the potential to transform almost everything around us. It can change the way humans interact with the world by making the objects around them "smart" — capable of constantly learning, adapting, and providing proactive assistance. The beginnings of this trend can already be seen in the new capabilities coming to smartphones (speech assistant, camera night mode) as well as the new class of “smart” devices such as smart watches, smart thermostats, and so on. However, these “smart” devices run much of the computation on the cloud (or a remote host) — costing them transmission power and response latency as well as causing potential privacy concerns. This limits their ability to provide a compelling user experience and realize the true potential of an “AI everywhere” world.

This workshop seeks to accelerate the transition towards a truly “smart” world where the AI capabilities permeate to all devices and sensors. The workshop will focus on how to distribute the AI capabilities across the whole system stack and co-design of edge device capabilities and AI algorithms. It will bring together researchers and practitioners with diverse backgrounds to cover the whole stack from application domains such as computer vision and speech, to the AI and machine learning algorithms that enable them, to the SoC/chip architecture that run them, and finally to the circuits, sensors, and memory technologies needed to build these devices.

Schedule:
https://research.fb.comprograms/on-device-intelligence-workshop/#Schedule

**Workshop Schedule Highlights**

**Morning Session:** Enabling new experiences on smart devices and agents

**Keynote Speaker:** [Blaise Aguera y Arcas](https://en.wikipedia.org/wiki/Blaise_Ag%C3%BCera_y_Arcas)

**Speaker Bio:**
Blaise leads an organization at Google AI working on both basic research and new products. Among the team’s public contributions are MobileNets, Federated Learning, Coral, and many Android and Pixel AI features. They also founded the Artists and Machine Intelligence program, and collaborate extensively with academic researchers in a variety of fields. Until 2014 Blaise was a Distinguished Engineer at Microsoft, where he worked in a variety of roles, from inventor to strategist, and led teams with strengths in interaction design, prototyping, machine vision, augmented reality, wearable computing and graphics. Blaise has given TED talks on Seadragon and Photosynth (2007, 2012), Bing Maps (2010), and machine creativity (2016). In 2008, he was awarded MIT’s TR35 prize.

**Afternoon Session:** Model, Software and Hardware co-design and optimization

**Keynote Speaker:** [Diana Marculescu](http://www.ece.utexas.edu/people/faculty/diana-marculescu)

**Speaker Bio:**
Diana Marculescu is Department Chair, Cockrell Family Chair for Engineering Leadership #5, and Professor, Motorola Regents Chair in Electrical and Computer Engineering #2, at the University of Texas at Austin. Before joining UT Austin in December 2019, she was the David Edward Schramm Professor of Electrical and Computer Engineering, the Founding Director of the College of Engineering Center for Faculty Success (2015-2019) and has served as Associate Department Head for Academic Affairs in Electrical and Computer Engineering (2014-2018), all at Carnegie Mellon University. She received the Dipl. Ing. degree in computer science from the Polytechnic University of Bucharest, Bucharest, Romania (1991), and the Ph.D. degree in computer engineering from the University of Southern California, Los Angeles, CA (1998). Her research interests include energy- and reliability-aware computing, hardware aware machine learning, and computing for sustainability and natural science applications. Diana was a recipient of the National Science Foundation Faculty Career Award (2000-2004), the ACM SIGDA Technical Leadership Award (2003), the Carnegie Institute of Technology George Tallman Ladd Research Award (2004), and several best paper awards. She was an IEEE Circuits and Systems Society Distinguished Lecturer (2004-2005) and the Chair of the Association for Computing Machinery (ACM) Special Interest Group on Design Automation (2005-2009). Diana chaired several conferences and symposia in her area and is currently an Associate Editor for IEEE Transactions on Computers. She was selected as an ELATE Fellow (2013-2014), and is the recipient of an Australian Research Council Future Fellowship (2013-2017), the Marie R. Pistilli Women in EDA Achievement Award (2014), and the Barbara Lazarus Award from Carnegie Mellon University (2018). Diana is a Fellow of both ACM and IEEE.

**Important Deadlines**
- Submission deadline: Jan 15, 2020
- Paper decision notification: Jan 27, 2020
- Presentation/Poster for accepted submissions: Feb 28, 2020

**Schedule**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00 AM</td>
<td>Enabling new experiences on smart devices and agents</td>
</tr>
<tr>
<td>10:30 AM</td>
<td>Imitation Learning from Observation by Prof Peter Stone</td>
</tr>
<tr>
<td>02:00 PM</td>
<td>Model, Software and Hardware Co-optimization</td>
</tr>
<tr>
<td>02:45 PM</td>
<td>How to Evaluate Deep Learning Accelerators by Prof. Vivienne Sze</td>
</tr>
<tr>
<td>03:15 PM</td>
<td>AutoML for on-device vision by Minxing Tan (Google)</td>
</tr>
</tbody>
</table>
This workshop will bring classical system architecture and design experts and AI/ML algorithmic experts together in one forum. The goal is to brainstorm about challenges in designing secure and resilient AI-centric systems in general, but with a special focus on autonomous systems (such as self-driving cars and industrial robots) - where safety and security are of paramount value.

The knowledge and expertise of classical mainframe and server architects who are experts in designing ultra-reliable and secure systems will be blended with domain experts in AI - particularly those with an established expertise in developing reliable and secure AI algorithms.

Detailed workshop information, abstract submission instructions, dates: https://sara-workshop.org

**Schedule**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session/Panel</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00 AM</td>
<td><strong>Introduction:</strong> Nandhini Chandramoorthy (IBM)</td>
</tr>
<tr>
<td>09:05 AM</td>
<td><strong>Keynote I:</strong> Dr. Thomas Rondeau (DARPA): Secure and Resilient - a DARPA View</td>
</tr>
<tr>
<td>09:50 AM</td>
<td><strong>Coffee Break + Discussion</strong></td>
</tr>
<tr>
<td>10:05 AM</td>
<td><strong>Energy-Efficient Circuits for Entropy Generation and Secure Encryption:</strong> Dr. Sanu Matthew (Intel Corp)</td>
</tr>
<tr>
<td>10:25 AM</td>
<td><strong>Feature Map Vulnerability Evaluation in CNNs:</strong> Abdurahman Mahmoud, Siva Kumar Sastry Hari, Christopher W. Fletcher, Charbel Sakr, Naresh Shanbag, Pavlo Molchanov, Michael B. Sullivan, Timothy Tsai, Stephen W. Keckler (UIUC and NVIDIA co-authors)</td>
</tr>
<tr>
<td>10:40 AM</td>
<td><strong>Reliable Intelligence in Unreliable Environment:</strong> Prof. Saibal Mukhopadhyay (Georgia Tech)</td>
</tr>
<tr>
<td>11:00 AM</td>
<td><strong>Towards Information Theoretic Adversarial Examples:</strong> Chia-Yi Hsu (NCHU), Pin-Yu Chen (IBM) and Chia-Mu Yu (NCHU)</td>
</tr>
<tr>
<td>11:15 AM</td>
<td><strong>Explaining Away Attacks Against Neural Networks:</strong> Sean Saito, Jin Wang (SAP Asia)</td>
</tr>
<tr>
<td>11:30 AM</td>
<td><strong>Poster Session + Discussion</strong></td>
</tr>
<tr>
<td>12:00 PM</td>
<td><strong>Lunch Break</strong></td>
</tr>
<tr>
<td>01:30 PM</td>
<td><strong>Poster Session + Discussion (Contd.)</strong></td>
</tr>
<tr>
<td>02:00 PM</td>
<td><strong>Keynote II:</strong> Prof. Xue Lin (Northeastern University): Towards Robust and Efficient Deep Learning Systems</td>
</tr>
<tr>
<td>02:45 PM</td>
<td><strong>MUTE: Data-Similarity Driven Multi-Hot Target Encoding for Neural Network Design:</strong> Mayoore Jaiswal, Bumsoo Kang, Jinho Lee, Minsik Cho (IBM)</td>
</tr>
<tr>
<td>03:00 PM</td>
<td><strong>WARDEN: Warranting Robustness Against Deception in Data Centers:</strong> Hazar Yueksel, Ramon Bertran, Alper Buyuktosunoglu (IBM)</td>
</tr>
<tr>
<td>03:15 PM</td>
<td><strong>Embedded Tutorial:</strong> Self-Progressing Robust Training:** Dr. Pin-Yu Chen (IBM Corp)</td>
</tr>
<tr>
<td>03:45 PM</td>
<td><strong>Coffee Break + Discussion</strong></td>
</tr>
<tr>
<td>04:00 PM</td>
<td><strong>Panel Discussion:</strong> Pin-Yu Chen (IBM), Akshay Deshpande (Soothsayer Analytics), Xue Lin (Northeastern University); Sean Saito (SAP, Asia); Moderators: Dr. Nandhini Chandramoorthy and Dr. Pradip Bose (IBM)</td>
</tr>
<tr>
<td>05:00 PM</td>
<td><strong>Closing Remarks:</strong> Organizers (IBM)</td>
</tr>
</tbody>
</table>
Abstract 4: Energy-Efficient Circuits for Entropy Generation and Secure Encryption: Dr. Sanu Matthew (Intel Corp) in SARA: Secure and Resilient Autonomy, 10:05 AM

Abstract: Physically Unclonable Functions (PUF) and True Random Number Generators (TRNG) are foundational security primitives underpinning the root of trust in computing platforms. Contradictory design strategies to harvest static and dynamic entropies typically necessitate independent PUF and TRNG circuits, adding to design cost. This tutorial describes a unified static and dynamic entropy generator leveraging a common entropy source for simultaneous PUF and TRNG operation. We will present self-calibration techniques to run-time segregate bitcells into PUF and TRNG candidates, along with entropy extraction techniques to maximize TRNG entropy while stabilizing PUF bits. Cryptographic circuits such as Advanced Encryption Standard (AES) are vulnerable to correlation power analysis (CPA) side-channel attacks (SCA), where an adversary monitors supply current signatures of a chip to decipher the value of embedded keys. This tutorial will also discuss the use of arithmetic/circuit countermeasures to minimize the correlation of the AES current to embedded keys, thereby improving the SCA resistance of the hardware by 1200x in both time and frequency-domains.

Bio: Bio: Sanu Mathew is a Senior Principal Engineer with the Circuits Research Labs at Intel Corporation, Hillsboro, Oregon, where he heads the security arithmetic circuits research group, responsible for developing special-purpose hardware accelerators for cryptography and security. He received his Ph.D. degree in Electrical and Computer Engineering from State University of New York at Buffalo in 1999. He holds 62 issued patents, has 20 patents pending and has published over 80 conference/journal papers. He is a Fellow of the IEEE.

Abstract 5: Feature Map Vulnerability Evaluation in CNNs: Abdulrahman Mahmoud, Siva Kumar Sastry Hari, Christopher W. Fletcher, Charbel Sakr, Naresh Shanbag, Pavlo Molchanov, Michael B. Sullivan, Timothy Tsai, Stephen W. Keckler (UIUC and NVIDIA co-authors) in SARA: Secure and Resilient Autonomy, 10:25 AM

ABSTRACT
As Convolutional Neural Networks (CNNs) are increasingly being employed in safety-critical applications, it is important that they behave reliably in the face of hardware errors. Transient hardware errors may percolate undesirable state during execution, resulting in software-manifested errors which can adversely affect high-level decision making. This talk will present HarDNN, a software-directed approach to identify vulnerable computations during a CNN inference and selectively protect them based on their propensity towards corrupting the inference output in the presence of a hardware error. We show that HarDNN can accurately estimate relative vulnerability of a feature map (fmap) in CNNs using a statistical error injection campaign, and explore heuristics for fast vulnerability assessment. Based on these results, we analyze the tradeoff between error coverage and computational overhead that the system designers can use to employ selective protection.

Abstract 6: Reliable Intelligence in Unreliable Environment: Prof. Saibal Mukhopadhyay (Georgia Tech) in SARA: Secure and Resilient Autonomy, 10:40 AM

Abstract: The artificial intelligence platforms are being increasingly deployed in safety-critical applications and autonomous systems, self-driving cars, robots, drones, to name a few. Unlike AI at the cloud environments, AI at these platforms need to perform reliably under changing environmental conditions and robust against different types of noise, while meeting stringent energy and time constraints. The reliability of AI platforms in unreliable environments is therefore a key challenge for deployment of AI in real-time safety-critical systems. This paper will present a broad perspective on how to design AI platforms to achieve this unique goal. First, we will present examples of AI architecture and algorithm that can assist in improving robustness against dynamic environment and noise, natural and adversarial. Next, we will discuss examples of how to make AI platforms robust against hardware induced noise and variation. The preceding discussions will focus on AI based on statistical machine learning models, including, deep learning. Finally, we will present a new generation of AI models that couple statistical learning with dynamical systems and neuro-inspired learning to enhance the reliability of AI models. The talk will conclude with future research opportunities and directions in this area.

Saibal Mukhopadhyay received his B. E. degree in Electronics and Telecommunication Engineering from Jadavpur University, Calcutta, India, in 2000. He received a Ph.D. degree in Electrical and Computer Engineering from Purdue University, West Lafayette, IN, in 2006. He was with the IBM T. J. Watson Research Center, Yorktown Heights, NY as a Research Staff Member. Since September 2007 he has been with the School of Electrical and Computer Engineering at the Georgia Institute of Technology, Atlanta, GA, where he is currently a Joseph M. Petit Professor of Electrical and Computer Engineering. His current research interests include neuromorphic computing and mixed-signal systems. Dr. Mukhopadhyay received the Office of Naval Research Young Investigator Award in 2012, the National Science Foundation CAREER Award in 2011, the IBM Faculty Partnership Award in 2009 and 2010, the SRC Inventor Recognition Award in 2008, the SRC Technical Excellence Award in 2005, and the IBM PhD Fellowship Award for years 2004-2005. He has received the IEEE Transactions on VLSI Systems (TVLSI) Best Paper Award in 2014, the IEEE Transactions on Component, Packaging, and Manufacturing Technology (TCPMT) Best Paper Award in 2014, the IEEE/ACM International Symposium on Low-power Electronic Design (ISLPED) Best Paper Award in 2014, the International Conference on Computer Design (ICCD) Best Paper Award in 2004, the IEEE Nano Student Paper Award in 2003, and multiple Best in Session Awards in SRC TECHCON in 2014 and 2005. He has authored or co-authored over 150 papers in refereed journals and conferences, and has been awarded six (6) U.S. patents. He is a Senior Member of IEEE.
to name a few. Besides advanced attack methods, this talk will discuss a concurrent adversarial training and model compression technique, which can achieve simultaneous model robustness and compactness for the deep learning applications in security-critical and resource-limited computing environment. The second part of the talk will introduce our hardware-aware deep neural network weight pruning method targeting the FPGA platforms and 3D convolutional neural networks for video recognition. Furthermore, the talk will discuss our new privacy-preserving weight pruning techniques.

Bio: Dr. Xue (Shelley) Lin is an assistant professor in the Department of Electrical and Computer Engineering at Northeastern University since 2017. She received her bachelor's degree in Microelectronics from Tsinghua University, China and her PhD degree from the Department of Electrical Engineering at University of Southern California in 2016. Her research interests include deep learning security and hardware acceleration, machine learning and computing in cyber-physical systems, high-performance and mobile cloud computing systems, and VLSI. Her research work has been recognized by several NSF awards and supported by Air Force Research Lab, Office of Naval Research, and Lawrence Livermore National Lab. She got the best paper award at ISVLSI 2014 and the top paper award at CLOUD 2014.

Abstract 13: MUTE: Data-Similarity Driven Multi-Hot Target Encoding for Neural Network Design: Mayooro Jaiswal, Bumsoo Kang, Jinho Lee, Minsik Cho (IBM) in SARA: Secure and Resilient Autonomy. 02:45 PM

ABSTRACT

Target encoding is an effective technique to deliver better performance for machine learning methods, but, existing approaches require significant increase in the learning capacity, thus demand higher computation power and more training data. In this paper, we present a novel and efficient target encoding scheme, MUTE to improve both generalizability and robustness of a target model by understanding the inter-class characteristics of a target dataset. By extracting the confusion-level between the target classes in a dataset, MUTE strategically optimizes the Hamming distances among target encoding. Such optimized target encoding offers higher classification strength for neural network models with negligible computation overhead and without increasing the model size.

Abstract 15: Embedded Tutorial: Self-Progressing Robust Training; Dr. Pin-Yu Chen (IBM Corp) in SARA: Secure and Resilient Autonomy, 03:15 PM

Abstract: Enhancing model robustness under new and even adversarial environments is a crucial milestone toward building trustworthy and reliable machine learning systems. Current robust training methods such as adversarial training explicitly specify an “attack” (e.g., Lp-norm bounded perturbation) to generate adversarial examples during model training in order to improve adversarial robustness. In this work, we take a different perspective and propose a new framework SPROUT, self-progressing robust training. During model training, SPROUT progressively adjusts training label distribution via our proposed parametrized label smoothing technique, making training free of attack generation and more scalable. We also motivate SPROUT using a general formulation based on vicinity risk minimization, which includes many robust training methods as special cases. Compared with state-of-the-art adversarial training methods (PGD and TRADES) under L-infinity-norm bounded attacks and various invariance tests, SPROUT consistently attains superior performance and is more scalable to large neural networks. Our results shed new light on scalable, effective and attack-independent robust training methods.

Bio: Dr. Pin-Yu Chen is currently a research staff member at IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA. He is also the chief scientist of RPI-IBM AI Research Collaboration and PI of ongoing MIT-IBM Watson AI Lab projects. Dr. Chen received his Ph.D. degree in electrical engineering and computer science and M.A. degree in Statistics from the University of Michigan, Ann Arbor, USA, in 2016. He received his M.S. degree in communication engineering from National Taiwan University, Taiwan, in 2011 and B.S. degree in electrical engineering and computer science (undergraduate honors program) from National Chiao Tung University, Taiwan, in 2009.

Dr. Chen’s recent research is on adversarial machine learning and robustness of neural networks. His long-term research vision is building trustworthy machine learning systems. He has published more than 20 papers on trustworthy machine learning at major AI and machine learning conferences, given tutorials at CVPR’20, ECCV’20, ICASSP’20, KDD’19 and Big Data’18, and co-organized several workshops for adversarial machine learning. His research interest also includes graph and network data analytics and their applications to data mining, machine learning, signal processing, and cyber security. He was the recipient of the Chia-Lun Lo Fellowship from the University of Michigan Ann Arbor. He received the NIPS 2017 Best Reviewer Award, and was also the recipient of the IEEE GLOBaCOM 2010 GOLD Best Paper Award. Dr. Chen is currently on the editorial board of PLOS ONE.

At IBM Research, Dr. Chen has co-invented more than 20 U.S. patents. In 2019, he received two Outstanding Research Accomplishments on research in adversarial robustness and trusted AI, and one Research Accomplishment on research in graph learning and analysis.

Abstract 18: Closing Remarks: Organizers (IBM) in SARA: Secure and Resilient Autonomy, 05:00 PM

Closing remarks and discussion on special journal issue.

CANCELED: Automated Machine Learning For Networks and Distributed Systems

behnazarzani Arzani, Bita Darvish Rouhani

Level 3 Room 10, Wed Mar 04, 09:00 AM

The first workshop on "Towards A Domain-Customized Automated Machine Learning Framework For Networks and Systems" aims at creating a coalition of researchers who aim to build an AutoML platform for network operators. The platform helps network operators bridge the expertise gap when using ML to solve challenging networking problems.

Researchers at this workshop will discuss how, as a community, can build a framework that: enables users to use ML to solve problems in networked systems without having in-depth ML expertise and that, similarly, enables ML experts to contribute to solving problems in networked systems without having expertise in these domains. We will discuss whether existing AutoML frameworks, as-is can be used by network operators? If not, what needs to change? Can domain
customization help? If yes, what are the components of a domain-customized AutoML framework and how are they different from traditional AutoML solutions? What are the important criteria that such a system needs to meet? What are the techniques we can use to build such a framework? What are the collaborations we can initiate across industry and academia to make headway on solving this problem?

Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00 AM</td>
<td>Introduction by Workshop Organizers</td>
</tr>
<tr>
<td>09:15 AM</td>
<td>The agony and the ecstasy of machine learning over the Internet</td>
</tr>
<tr>
<td>09:45 AM</td>
<td>Panelists: Victor Bahl (Microsoft Research), Nicolo Fusi (Microsoft Research Cambridge), Ranjita Bhagwan (Microsoft Research India)</td>
</tr>
<tr>
<td>11:00 AM</td>
<td>Panel, Part 2</td>
</tr>
<tr>
<td>12:00 PM</td>
<td>Lunch</td>
</tr>
<tr>
<td>01:30 PM</td>
<td>Accepted Talks</td>
</tr>
<tr>
<td>03:00 PM</td>
<td>Invited talks</td>
</tr>
<tr>
<td>04:00 PM</td>
<td>Round table discussions</td>
</tr>
</tbody>
</table>

Abstracts (3):

Abstract 2: The agony and the ecstasy of machine learning over the Internet in CANCELED: Automated Machine Learning For Networks and Distributed Systems, 09:15 AM

Many networking settings ask us, or really our computers, to make tough decisions from partial information: congestion control, traffic engineering, provisioning, channel characterization, scheduling, query planning, spam filtering, video streaming, predicting real-life indicators from crowdsourced information, etc. This suggests a natural setting for machine learning, which has shown great success in adjacent areas of computer science. And yet -- and yet! The Internet has turned out to be a particularly challenging setting for ML. We don't know how to simulate it, which makes it challenging to learn reliable control algorithms. It's hard to measure loss functions on a distributed network where each node receives only partial information. And we have a hard time learning algorithms that are robust to adversarial input. I'll present findings from two multi-year deployments of ML over the Internet, for congestion control and video streaming, and discuss where I think there's cause for optimism and caution.

Abstract 6: Accepted Talks in CANCELED: Automated Machine Learning For Networks and Distributed Systems, 01:30 PM

Taurus: An Intelligent Data Plane
Tushar Swamy (Stanford University), Alexander Rucker (Stanford University), Muhammad Shahbaz (Stanford University), and Kunle Olukotun (Stanford University)

Abstract 7: Invited talks in CANCELED: Automated Machine Learning For Networks and Distributed Systems, 03:00 PM

Daniel Berger (Experience with developing ML for distributed caching systems)

Organizers

Panelists: Victor Bahl (Microsoft Research), Nicolo Fusi (Microsoft Research Cambridge), Ranjita Bhagwan (Microsoft Research India)

MLOps Systems

Debo Dutta, Matei Zaharia, Ce Zhang

Level 3 Room 5, Wed Mar 04, 09:00 AM

Due to the complexity in putting ML into production, the actual machine learning capability is a small part of a complex system and its lifecycle. This new evolving field is known as MLOps. Informally MLOps typically refers to the collaboration between data scientists and operations engineers (e.g. SRE) to manage the lifecycle of ML within an organization. This space is new and has yet to be explored from a research perspective.

In this workshop we aim to cover research problems in MLOps, including the systems and ML challenges involved in this process. We will also cover the software engineering questions including specification, testing and verification of ML software systems. We will bring together a wide variety of experts from both industry and academia, covering persona ranging from data scientists to machine learning engineers.

For more information please visit the webpage
http://mlops-systems.github.io

Benchmarking Machine Learning Workloads on Emerging Hardware

Tom St John, Murali Emani

Level 3 Room 6, Wed Mar 04, 09:00 AM

With evolving system architectures, hardware and software stacks, diverse machine learning (ML) workloads, and data, it is important to understand how these components interact with each other. Well-defined benchmarking procedures help evaluate and reason the performance gains with ML workload-to-system mappings. We welcome all novel submissions in benchmarking machine learning workloads from all disciplines, such as image and speech recognition, language processing, drug discovery, simulations, and scientific applications. Key problems that we seek to address are: (i) which representative ML benchmarks cater to workloads seen in industry, national labs, and interdisciplinary sciences; (ii) how to characterize the ML workloads based on their interaction with hardware; (iii) which novel aspects of hardware, such as heterogeneity in compute, memory, and networking, will drive their adoption; (iv) performance modeling and projections to next-generation hardware. Along with selected publications, the workshop program will
also have experts in these research areas presenting their recent work and potential directions to pursue.

Call for Papers can be found here:
https://memani1.github.io/challenge20/

Paper Submission Deadline: January 15, 2020
Author Notification: January 27, 2020
Camera-Ready Papers Due: February 21, 2020

Abstracts (2):

Abstract 4: Morning Paper Session in Benchmarking Machine Learning Workloads on Emerging Hardware, 10:30 AM

Precious: Resource-Demand Estimation for Embedded Neural Network Accelerators - Stefan Ralf (FAU Erlangen-Nürnberg), Benedict Herzog (FAU Erlangen-Nürnberg), Judith Hemp (FAU Erlangen-Nürnberg), Timo Höning (FAU Erlangen-Nürnberg), Wolfgang Schröder-Preikschat (FAU Erlangen-Nürnberg)

Benchmarking Machine Learning Workloads in Structural Bioinformatics Applications - Heng Ma (Argonne National Laboratory), Austin Clyde (Argonne National Laboratory), Venkatram Vishwanath (Argonne National Laboratory), Debsindhu Bhowmik (Oak Ridge National Laboratory), Arvind Ramanathan (Argonne National Laboratory), Shantenu Jha (Rutgers University, Brookhaven National Laboratory)

Benchmarking Alibaba Deep Learning Applications Using AI Matrix - Wei Zhang (Alibaba Group), Wei Wei (Alibaba Group), Lingjie Xu (Alibaba Group), Lingling Jin (Alibaba Group)

Abstract 8: Afternoon Paper Session in Benchmarking Machine Learning Workloads on Emerging Hardware, 02:45 PM

Deep Learning Workload Performance Auto-Optimizer - Connie Yingyu Miao (Intel Corporation), Andrew Yang (Intel Corporation), Michael Anderson (Intel Corporation)

Challenges with Evaluating ML Solutions in Data Centers - Shobhit Kanaujia (Facebook), Wenyan Fu (Facebook), Abhishek Dhanotia (Facebook)

Benchmarking TinyML Systems: Challenges and Direction - Colby Banbury (Harvard University), Vijay Janapa Reddi (Harvard University), Will Fu (Harvard University), Max Lam (Harvard University), Amin Fazel (Samsung Semiconductor Inc.), Jeremy Holleman (Syntiant, University of North Carolina Charlotte), Xinyuan Huang (Cisco Systems), Robert Hurtado (HurtadoTechnology Inc.), David Kanter (Real World Insights), Anton Lokhmotov (dividiti), David Patterson (University of California Berkeley, Google), Danilo Pau (STMicroelectronics), Jeff Sieracki (Reality AI), Jae-Sun Seo (Arizona State University), Urmish Thakkar (Arm), Marian Verhelst (Umicor)

Resource-Constrained Machine Learning (ReCoML 2020)

Yaniv Ben Itzhak, Nina Narodytska, Christopher Aberger

Level 3 Room B, Wed Mar 4, 09:00 AM

** Notice that you need to register for the conference to be able to attend the workshop **

The workshop will cover broad aspects that are related to ML over resource-constrained environments, such as Internet-of-Things (IoT) devices, and edge-computing. Resource-constrained ML is challenging due to several reasons: First, current ML models usually have high resource requirements in terms of CPU, memory and I/O. Naive solutions that reduce these resource consumption would result in significant ML performance degradation. Therefore, new ML models and frameworks are required in order to employ ML with reasonable ML performance over resource-constrained environments. Second, resource-constrained environments, such as edge computing and IoT, are usually being used for real-time applications. Hence, the model serving is a critical issue, such that an ML model should respond quickly and accurately while being employed over limited resources. The workshop will specifically include the following topics: model/hardware architectures, models
Topics of interest include, but are not limited to:

- Compression of deep ML model architectures
- Quantized and low-precision neural networks
- Optimization of ML model architectures for resource-constrained environments
- Hardware accelerators for deep ML models
- Explainability of ML models in the context of resource-constrained environments
- ML deployments over resource-constrained environments, e.g. Internet-of-Things (IoT) devices and edge-computing.

**Reviewing process:** All submissions should include the author’s names and their affiliations. The authors are allowed to post their paper on arXiv or other public forums.

Key dates related to the reviewing process are given below:

**Paper submission deadline:** January 15, 2020 AoE (at midnight anywhere on earth)

**Decision notification:** January 27, 2020

We invite research contributions in different formats:

- Original research papers (up to 6 pages, not including references)
- Position, opinion papers and extended abstracts (up to 4 pages, not including references)

**Submission link:** [link](#)

**Dual submission policy:** We will not accept any paper which, at the time of submission, is under review for another workshop or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published.

**Proceedings:** Accepted papers will be published in the form of online proceedings.

**Submission format:** To prepare your submission to ReCoML 2020, please use the LaTeX style files provided at [SML2020style.tar.gz](#). Submitted papers will be in a 2-column format, each reference must explicitly list all the authors of the paper.

**Organizing Committee**

Yaniv Ben-Itzhak, VMware Research, ybenitzhak (at) vmware (dot) com

Nina Narodytska, VMware Research, nnarodytska (at) vmware (dot) com

Christopher R. Aberger, Stanford and SambaNova Systems, christopher.aberger (at) sambanovasystems (dot) ai

**Schedule**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:25 AM</td>
<td>Welcome</td>
</tr>
<tr>
<td>10:30 AM</td>
<td>Shared Clusters for Machine Learning: Through the looking glass, by Prof. Shivaram Venkataraman, University of Wisconsin</td>
</tr>
<tr>
<td>11:15 AM</td>
<td>QuaRL: Quantized Reinforcement Learning</td>
</tr>
<tr>
<td>11:30 AM</td>
<td>Optimizing Sparse Matrix Operations for Deep Learning</td>
</tr>
<tr>
<td>11:45 AM</td>
<td>Energy-Aware DNN Graph Optimization</td>
</tr>
<tr>
<td>12:00 PM</td>
<td>Lunch</td>
</tr>
<tr>
<td>02:00 PM</td>
<td>Low-Precision Arithmetic in Machine Learning Systems, by Prof. Christopher De Sa, Cornell</td>
</tr>
<tr>
<td>02:45 PM</td>
<td>Efficient Memory Management for Deep Neural Net Inference</td>
</tr>
<tr>
<td>03:00 PM</td>
<td>Once for All: Train One Network and Specialize it for Efficient Deployment</td>
</tr>
<tr>
<td>03:15 PM</td>
<td>GReTA: Hardware Optimized Graph Processing For GNNs</td>
</tr>
<tr>
<td>03:30 PM</td>
<td>Afternoon Break</td>
</tr>
<tr>
<td>04:00 PM</td>
<td>Optimizing JPEG Quantization for Classification Networks</td>
</tr>
<tr>
<td>04:15 PM</td>
<td>Conditional Neural Architecture Search</td>
</tr>
<tr>
<td>04:30 PM</td>
<td>Transfer Learning with Fine-grained Sparse Networks: From Efficient Network Perspective</td>
</tr>
<tr>
<td>04:45 PM</td>
<td>Closing remarks</td>
</tr>
</tbody>
</table>

**Abstracts (2):**

**Abstract 2:** Shared Clusters for Machine Learning: Through the looking glass, by Prof. Shivaram Venkataraman, University of Wisconsin in Resource-Constrained Machine Learning (ReCoML 2020), 10:30 AM

With recent advances in machine learning, large enterprises incorporate machine learning models across a number of products. To facilitate training of these models, enterprises use shared, multi-tenant cluster of machines equipped with accelerators like GPUs. Similar to data analytics clusters, operators aim to achieve high resource utilization while providing resource isolation and fair sharing across users. In this talk we will first present characterization of machine learning workloads from a multi-tenant GPU cluster at Microsoft. We then present how various aspects of these workloads such as gang scheduling and locality constraints affect resource utilization and efficiency. Based on this analysis we discuss research efforts to improve efficiency and utilization both for individual jobs and across the cluster.
Abstract 7: Low-Precision Arithmetic in Machine Learning Systems, by Prof. Christopher De Sa, Cornell in Resource-Constrained Machine Learning (ReCoML 2020), 02:00 PM

Much of the recent advancement in machine learning has been driven by the capability of machine learning systems to process and learn from very large data sets using very complicated models. Continuing to scale data up in this way presents a computational challenge, as power, memory, and time are all factors that limit performance. One popular approach to address these issues is low-precision arithmetic in which a lower-precision number is used to improve these system metrics—although possibly at the cost of some accuracy. In this talk, I will discuss some recent methods from my lab that use numerical precision for ML tasks, while the same time trying to understand its effects theoretically.

Software-Hardware Codesign for Machine Learning Workloads

Ritwik Gupta, John Wohlbier, Tze Meng Low, Jeffrey Vetter, Natalia Vassilieva

Level 3 Room 9, Wed Mar 04, 09:00 AM

Machine learning development workflows today involve the siloed design and optimization of task-specific software for a limited number of fixed hardware options. As a result, hardware and software are seen as individual components where the impact of either SW or HW on each other cannot be optimized or assessed jointly. This abstraction leads to computationally inefficient machine learning workloads.

Recently, both software and hardware have taken steps to become more domain specific. Machine learning focused software libraries provide operations and abstractions limited to workload-relevant use cases. Hardware makers have started manufacturing workload-relevant chips in the form of FPGAs, ASICs, and DLAs. However, these efforts are still largely independent of each other, resulting in inefficiencies and less-than-ideal workload performances.

Ideally, hardware and software would be codesigned for a specific ML workload, but investing in a particular hardware design is costly, especially in the face of the rapidly evolving state of ML. This workshop is soliciting extended abstracts that seek to bridge the gap between software and hardware in the areas of model design, model abstractions, model primitives, workload compression, hardware design, hardware optimization for power, data flow optimization, and compiler technologies.

Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00 AM</td>
<td>Welcome, Introduction, Logistics</td>
</tr>
<tr>
<td>08:10 AM</td>
<td>DARPA</td>
</tr>
<tr>
<td>08:35 AM</td>
<td>SambaNova</td>
</tr>
<tr>
<td>09:00 AM</td>
<td>Groq</td>
</tr>
<tr>
<td>09:25 AM</td>
<td>Graphcore</td>
</tr>
<tr>
<td>09:50 AM</td>
<td>Break</td>
</tr>
<tr>
<td>10:05 AM</td>
<td>Cerebras</td>
</tr>
</tbody>
</table>

Abstracts (12):

Abstract 2: DARPA in Software-Hardware Codesign for Machine Learning Workloads, 08:10 AM

Dr. Thomas Rondeau - Program Manager - DARPA

Title:

Abstract: AWAITING DARPA RELEASE

Keywords: software defined hardware, domain specific SoC, reconfigurable hardware, many core heterogeneous specialized accelerators

Bio: Tom Rondeau is a program manager in DARPA's Microsystems Technology Office with a focus on adaptive and reconfigurable radios, improving the development cycle for new signal-processing techniques, and exploring new approaches and applications with the electromagnetic spectrum. Prior to joining DARPA, Tom was the maintainer and lead developer of the GNU Radio project, a visiting researcher with the University of Pennsylvania, and an Adjunct with the IDA Center for Communications Research in Princeton, NJ.

Abstract 3: SambaNova in Software-Hardware Codesign for Machine Learning Workloads, 08:35 AM

Dr. Christopher Aberger - Director, Software Engineering - SambaNova Systems

Title:

Abstract: In many applications traditional software development is being replaced by machine learning generated models resulting in accuracy improvements and deployment advantages. This fundamental shift in how we develop software is known as Software 2.0. The continued success of Software 2.0 will require efficient and flexible computer hardware optimized for the dataflow computational graphs at the core of machine learning. In this talk, we will discuss the design of high-performance dataflow computer architectures for machine learning. Our vertically integrated approach to machine learning performance combines new machine learning algorithms, new domain-specific languages, advanced compilation technology and software-defined...
Abstract 4: Groq in Software-Hardware Codesign for Machine Learning Workloads, 09:00 AM

Dr. Dennis Abts - Chief Architect - Groq

Title: From Supercomputers to Superchips: Deep Learning One PataOp at a Time

Abstract:

Keywords: tensor streaming architecture, inference, software defined hardware, compiler technology

Bio: Dennis is the Chief Architect at Groq, and is an expert in scalable vector architectures for high performance computing. Previously at Google, he worked on datacenter network topologies for energy-proportional networking and Cray where he was a Sr. Principal Architect on several Top500 massively-parallel supercomputers. Dennis has published over 20 technical papers in areas of memory systems, interconnection networks, and fault-tolerant systems. He holds over two dozen patents spanning 20+ years of experience at Cray and Google. Dennis holds a Ph.D. in Computer Science from the University of Minnesota and is a Senior Member of IEEE and ACM Computer Society.

Abstract 5: Graphcore in Software-Hardware Codesign for Machine Learning Workloads, 09:25 AM

Matt Fyles - VP Software - Graphcore

Title: Compiling For Distributed Memory Architectures

Abstract: The Graphcore Intelligence Processing Unit (IPU) is designed for targeting machine learning workloads and supporting the scaling of applications across multiple devices. The IPU architecture is based around massively parallel distributed processing where applications are mapped over thousands of processor cores and operate using a Bulk Synchronous Parallel (BSP) execution model which separates computation from communication. In order to achieve performance from applications mapped onto the IPU the software tool chain has to deal with the complex task of partitioning machine learning computational graphs. In this presentation we discuss how we take a machine learning application and through our software tools partition and schedule the work across the IPU. We also discuss the hardware / software trade-offs that were made to build a processor to execute these workloads.

Keywords: IPU, ML computational graph, bulk synchronous parallel, training, inference

Bio: Matt Fyles is a computer scientist with over 20 years experience in the design, development, delivery and support of software and hardware for the microprocessor market, spanning a wide range of applications from consumer electronics to high performance computing, with a particular focus on parallel processors. He began his career at STMicroelectronics, Europe’s largest semi-conductor company, followed by SuperH, Clearspeed and XMOS. He is currently Vice President of Software at Graphcore, a Bristol-based artificial intelligence hardware and software company. Matt is a graduate of Computer Science from the University of Exeter.

Abstract 7: Cerebras in Software-Hardware Codesign for Machine Learning Workloads, 10:05 AM

Dr. Natalia Vassilieva - Technical Product Manager - Cerebras Systems

Title: Accelerating Deep Learning with a purpose-built solution: the Cerebras approach

Abstract: The new era of chip specialization for deep learning is here. Traditional approaches to computing can no longer meet the computational and power requirements of this workload, arguably the most important of our generation. What is the right processor for deep learning? To answer this question, this talk will provide an overview of deep neural nets, discuss computational requirements of different types of models and limitations of existing hardware architectures and scale-out approaches. Then we will discuss Cerebras’ approach to meet computational requirements of deep learning with the Cerebras Wafer Scale Engine (WSE) -- the largest computer chip in the world, and the Cerebras Software Platform, co-designed with the WSE. The WSE provides cluster-scale resources on a single chip with full utilization for tensors of any shape -- fat, square and thin, dense and sparse -- enabling researchers to explore novel network architectures and optimization techniques at any batch sizes. Finally, we will discuss potential co-design ideas for new neural net models and learning methods for the WSE.

Keywords: WSE, training, inference, dataflow computational graph

Bio: Natalia Vassilieva is a Technical Product Manager at Cerebras Systems, a computer systems company dedicated to accelerating deep learning. Her focus is machine learning and artificial intelligence, analytics, and application-driven software-hardware optimization and co-design. Most recently before joining Cerebras Natalia has been a Sr. Research Manager at Hewlett Packard Labs, where she led the Software and AI group and served as the head of HP Labs Russia from 2011 till 2015. Prior to HPE, she was an Associate Professor at St. Petersburg State University and worked as a software engineer for different IT companies. Natalia holds a PhD in computer science from St. Petersburg State University.

Abstract 8: Oak Ridge National Laboratory in Software-Hardware Codesign for Machine Learning Workloads, 10:30 AM

Dr. Jeffrey Vetter - Future Technologies Group Leader - Oak Ridge National Laboratory

Title:

Abstract:

Bio:
Bio: Tze Meng Low is an Assistant Research Professor with the Department of Electrical and Computer Engineering at Carnegie Mellon University. He graduated from the University of Texas at Austin with an M.S.(C.S) in 2004, and a Ph.D. in Computer Science in 2013. His research focuses on the systematic derivation and implementation of high-performance algorithms through the use of formal methods and analytical models. His goal is to achieve performance portability across both architectures and domains by understanding and capturing the interaction between software algorithms and hardware features through analytical models so as to build better code-generators, and/or software libraries for emerging domains and architectures.

Abstract: Recent advances in machine learning (ML) have depended on the continued progress of hardware computing platforms. Future advances will depend even more on the synergistic progress of hardware and software. This is the case particularly for embedded ML applications, where developers must meet performance requirements under tighter resource constraints. The emerging open-source hardware community can play a unique role in supporting embedded ML research. ESP is an open-source research platform to design and program heterogeneous systems-on-chip. With the design automation capabilities of ESP, application developers can synthesize hardware accelerators from models specified in common ML frameworks, integrate these accelerators in a complete system-on-chip, and quickly obtain FPGA-based prototypes to evaluate their design by running embedded ML applications.

Keywords: SoC design, ML frameworks, FPGA

Bio: Luca Carloni is Professor of Computer Science at Columbia University in the City of New York. He holds a Laurea Degree Summa cum Laude in Electronics Engineering from the University of Bologna, Italy, and the MS and PhD degrees in Electrical Engineering and Computer Sciences from the University of California, Berkeley. His research interests include methodologies and tools for system-on-chip platforms with emphasis on heterogeneous computing, intellectual property reuse, design of networks-on-chip, embedded software, and distributed embedded systems. He coauthored over one hundred and fifty refereed papers and is the holder of two patents. Luca received the Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2006, was selected as an Alfred P. Sloan Research Fellow in 2008, and received the ONR Young Investigator Award and the IEEE CEDA Early Career Award in 2010 and 2012, respectively. In 2013 Luca served as general chair of Embedded Systems Week (ESWeek), the premier event covering all aspects of embedded systems and software. Luca is an IEEE Fellow.

Abstract: AI scientists are moving from research (training) using high price, high power, large form factor HPCs to productization (inference). AI inference requires orders of magnitude more horsepower while keeping the price, power, latency, form factor intact, Xilinx adaptable devices are ideal for that. However, the biggest challenge has been the programming model where it required developers to be hardware savvy. In this talk, we will introduce the newly released development environment called Vitis AI, which allows users to directly take their TensorFlow trained models and target Xilinx devices from edge to cloud. Vitis AI consists of a suite of familiar tools for AI scientists: quantizer, pruner, compiler, profiler, runtime, and pre-optimized Deep Learning Processing Units (DPU).

Keywords: ML toolkit, FPGA, DPU

Bio: Nick Ni is the director of product marketing, AI and software at Xilinx. His team’s responsibilities include business development, go-to-market plans, ecosystem development, and outbound marketing for Xilinx’s artificial intelligence products and software/hardware development tools. Ni joined Xilinx in 2014. Before Xilinx, he held multiple roles in R&D and applications at ATI, AMD, Qualcomm, and Intel, focusing on embedded systems design and high-level synthesis. Ni earned a master’s degree in Computer Engineering from the University of Toronto and holds over 10 patents and publications.

Abstract: We propose to show a Deep Learning Inference toolkit (OpenVINO), which provides a common API for inference independent of the underlying compute hardware. The inference engine can operate on CPU or be accelerated with GPU, VPU or FPGA. We will further look into details of an OpenCL based Deep Learning Accelerator running on FPGA and how this is integrated into the software flow. We will conclude
with a brief discussion of the potential use of the oneAPI unified
programming model could be used for future developments of such
hardware agnostic accelerators.

Keywords: Inference, ML toolkit, CPU, GPU, VPU, FPGA

Bio: James Moawad is a Technical Solution Specialist with Intel's
Programmable Solutions Group specializing in compute acceleration
using Field Programmable Gate Arrays (FPGA). He holds a B.S. in
Electrical Engineering from the University of Illinois at
Urbana-Champaign and a M.S. in Electrical and Computer Engineering
from Georgia Institute of Technology with a focus on processor
architecture. He designed telecommunication systems at Bell
Laboratories / Lucent Technologies from 1999 to 2006 utilizing FPGAs
and multi-processor arrays. Since 2006, he has worked as a Field
Application Engineer helping customers architect systems with FPGA,
embedded processors, DSP and various memory solutions including
DRAM, solid state drives and high bandwidth memory (HBM).

Abstract 17: Arm in Software-Hardware Codesign for Machine
Learning Workloads, 04:00 PM

Dr. Kshitij Sudan - Principle Solutions Architect - Arm

Title:

Abstract: Machine learning processing gets a lot of attention due to novel
hardware accelerators being developed to speed-up emerging
use-cases. The large and rapidly evolving accelerator space for ML
processing however is eclipsed in reality by the amount of ML processing
that happens on general purpose CPUs. Some estimates rate >80% of
ML inference to occur on general-purpose CPUs. The driving factors for
on-CPU processing are three fold: 1) Ease of programming, 2) Integration of ML analysis output with business applications, 3) Duty-cycle of ML workloads. In this talk we will first outline the use-cases that are well served by on-CPU ML workload execution followed by how Arm is working to enable more efficient use of general-purpose Arm
CPUs for edge-to-cloud processing of ML workloads. Efficient processing
requires both hardware and software features to be co-developed –
especially since ML algorithms are rapidly evolving. Arm is leveraging
this co-design philosophy along with its traditional strength in energy
efficient design to make on-CPU ML processing pervasive and
easy-to-use.

Keywords: inference, CPU

Bio: Dr. Kshitij Sudan is a Principal Solution Architect in the Infrastructure
Business Unit at Arm where he helps build solutions to address market
and customer needs. A solution could either be a single piece of Arm IP
or a whole platform offering consisting of Arm IP and enabling
open-source software stack. His current areas of focus include
smart-offload (like SmartNICs), platform security, video encoding, and
efficient ML/AI processing. He received his Ph.D. from the University of
Utah where his research focused on DRAM-based memory systems. He
has been granted two US patents and has multiple applications in the
pipeline.