Guest Editors' Introduction: Robust Resource-Constrained Machine Learning

Theocharis Theocharides
University of Cyprus

Muhammad Shafique
Technische Universität Wien

Jungwook Choi
Hanyang University

Onur Mutlu
ETH Zürich

Machine learning (ML) is nowadays embedded in several computing devices, consumer electronics, and cyber–physical systems. Smart sensors are deployed everywhere, in applications such as wearables and perceptual computing devices, and intelligent algorithms power our connected world. These devices collect and aggregate volumes of data, and in doing so, they augment our society in multiple ways; from healthcare, to social networks, to consumer electronics, and many more. To process these immense volumes of data, ML is emerging as the *de facto* analysis tool that powers several aspects of our Big Data society. Applications spanning from infrastructure (smart cities, intelligent transportation systems, smart grids, and to name a few), to social networks and content delivery, to e-commerce and smart factories, and emerging concepts such as self-driving cars and autonomous robots, are powered by ML technologies. These emerging systems require real-time inference and decision support; such scenarios, therefore, may use customized hardware accelerators, are typically bound by limited resources, and are restricted to limited connectivity and bandwidth. Thus, near-sensor computation and near-sensor intelligence have started emerging as necessities to continue supporting the paradigm shift of our connected world. The need for real-time intelligent data analytics (especially in the era of Big Data) for decision support near the data acquisition points emphasizes the need for revolutionizing the way we design, build, test, and verify processors, accelerators, and systems that facilitate ML (and deep learning, in particular) implemented in resource-constrained environments for use at the edge and the fog. As such, traditional von Neumann architectures are no longer sufficient and suitable, primarily because of limitations in both performance and energy efficiency caused especially by large amounts of data movement. Furthermore, due to the connected nature of such systems, security and reliability are also critically important. Robustness, therefore, in the form of reliability and operational capability in the presence of faults, whether malicious or accidental, is a critical need for such systems. Moreover, the operating nature of these systems relies on input data that is characterized by the four "V's": velocity (speed of data generation), variability (variable forms and types), veracity (unreliable and unpredictable), and volume (i.e., large amounts of data). Thus, the robustness of such systems needs to consider this issue as well. Furthermore, robustness in terms of security, and in terms of reliability to hardware and software faults, in

Digital Object Identifier 10.1109/MDAT.2020.2971201
Date of current version: 20 April 2020.

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

particular, besides their importance when it comes to safety-critical applications, is also a positive factor in building trustworthiness toward these disrupting technologies from our society. To achieve this envisioned robustness, we need to refocus on problems such as design, verification, architecture, scheduling and allocation policies, optimization, and many more, for determining the most efficient, secure, and reliable way of implementing these novel applications within a robust, resource-constrained system, which may or may not be connected. This special issue, therefore, addresses a key aspect of fog and edge-based ML algorithms; robustness (as defined above) under resource-constraint scenarios. The special issue presents emerging works in how we design robust systems, both in terms of reliability as well as fault tolerance and security, while operating with a limited number of resources, and possibly in the presence of harsh environments that may eliminate connectivity and pollute the input data.

This special issue features two keynote contributions, academic and an industrial one, offering respective viewpoints and discussing state-of-the-art issues in training for robustness and on emerging architectures and technologies such as resistive RAM. In particular, the first Keynote article by Seshia et al. titled "Semantic Adversarial Deep Learning," accounts for the semantics, context, and specifications of a complete system with ML components in resource-constrained environments, focusing on adversarial training for the robustness of deep neural networks (DNNs). The second Keynote article by Rasch et al., titled "Training Large-Scale Artificial Neural Networks on Simulated Resistive Crossbar Arrays," proposes a novel simulation framework for resistive crossbar arrays. Resistive crossbar arrays are promising options for accelerating enormous computation necessary for training modern DNNs, but verification of such systems has not been scaled up to realistic size problems. Thus, the Keynote article proposes a simulator that is capable of exploring design constraints on large-scale problems and enabling designers to devise algorithmic measures to pave the way for robust resistive crossbar-based DNN training accelerators. A Survey paper by Shafique et al. titled "Robust Machine Learning Systems: Challenges, Current Trends, Perspectives, and the Road Ahead," is also included that taxonomizes the challenges and opportunities in robust resource-constrained ML systems, summarizing the prominent vulnerabilities of such systems, and that highlights successful defenses and mitigation techniques against these vulnerabilities, both during the training phase and during the inference stage. The survey paper discusses the implications of a resource-constrained design on the reliability and security of the system, identifies verification methodologies to ensure correct system behavior, and describes open research challenges for building secure and reliable ML systems.

The special issue additionally features six contributed articles covering a broad range of issues and challenges. The first article, titled "SSCNets: Robustifying DNNs Using Secure Selective Convolutional Filters" by Ali et al., introduces a novel technique that is based on selective secure convolutional approach during training that increases robustness in DNNs by allowing the trained network model to learn data distribution, which is based on image features, namely the edges. The second article, titled "Adaptive Neural Network Architectures for Power-Aware Inference" by Anderson et al., proposes an adaptive approach in boosting the performance of a neural network during the inference stage, based on the available power, without completely having to reconfigure the neural network parameters. The third article, titled "Are CNNs Reliable Enough for Critical Applications?—An Exploratory Study" by Neggaz et al., investigates the impact of reliability issues on the underlying architectures, which facilitate convolutional neural networks. Through experimental fault injection approaches, the article investigates the impact of faults across various layers and discusses the vulnerability of the convolutional neural network and the host architecture. The fourth article, titled "Impact of Memory Voltage Scaling on Accuracy and Resilience of Deep-Learning-Based Edge Devices" by Denkinger et al., investigates how energy-reducing techniques such as quantization (which limits the size and number of accesses in memories) and voltage scaling, impact the overall robustness in edge devices. Energy savings and ways to increase them while maintaining reliable operation is also the theme of the fifth article, titled "Enabling Timing Error Resilience for Low-Power Systolic-Array-Based Deep-Learning Accelerators" by Zhang et al. The authors propose a mechanism that facilitates aggressive voltage scaling while at the same time coping with timing errors as well as process variation errors. The sixth and last article, titled "Backdoor Suppression in Neural Networks

Using Input Fuzzing and Majority Voting" by Sarkar et al., completes this special issue with another major challenge, particularly the security of neural inference. In particular, it addresses the case where inference is needed at the edge while training is typically done at the cloud. Thus, the trained model and training data are interchanged between the edge and the cloud creating a significant security hole such as inclusion of backdoors. This article discusses an approach where a trained model can still operate as expected, irrespective of the presence of such backdoors. ■

Theocharis (Theo) Theocharides has been an Associate Professor at the Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus, since 2006, where he directs the Embedded and Application-Specific Systems-on-Chip Laboratory. He has also been a Faculty Member of the KIOS Research and Innovation Center of Excellence, University of Cyprus, since the Center's inception in 2008 and the Research Director since the Center's upgrade to a European Research Center of Excellence in 2017. His research encapsulates the design, development, implementation, and deployment of low-power and reliable on-chip application-specific architectures, low-power VLSI design, real-time embedded systems design, and exploration of energy-reliability tradeoffs for systems on chip and embedded systems. His focus lies on acceleration of computer vision and artificial intelligence algorithms in hardware, geared toward edge computing, and in utilizing reconfigurable hardware toward self-aware, evolvable, and robust intelligent edge computing systems. Theocharides has a PhD in computer engineering from Penn State University, State College, PA, where he was working in the areas of low-power computer architectures and reliable system design with emphasis on computer vision and machine learning applications. He is a Senior Member of the IEEE and a member of IEEE CEDA and the ACM.

Muhammad Shafique has been a Full Professor at Computer Architecture and Robust Energy-Efficient Technologies (CARE-Tech.), Institute of Computer Engineering, Technische Universität Wien, Vienna, Austria, since November 2016. His research interests are in computer architecture, power-/energy-efficient systems, robust computing, hardware security, brain-inspired computing trends like neuromorphic and approximate computing, hardware and system-level design for machine learning and artificial intelligence (AI), emerging technologies, nanosystems, FPGAs, MPSoCs, and embedded systems. His research has a special focus on cross-layer modeling, design, and optimization of computing and memory systems, and their deployment in use cases from Internet-of-Things (IoT), cyber-physical systems (CPSs), and ICT for development (ICT4D) domains. Shafique has a PhD in computer science from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany (2011). He is a Senior Member of the IEEE and the IEEE Signal Processing Society (SPS), and a member of the ACM, SIGARCH, SIGDA, SIGBED, and HIPEAC.

Jungwook Choi is currently an Assistant Professor at Hanyang University, Seoul, South Korea. His research interests include high performance, energy efficient, and reliable implementation of machine learning and deep learning algorithms. Choi has a PhD in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Champaign, IL.

Onur Mutlu is a Professor of computer science at ETH Zürich, Zürich, Switzerland. He is also a Faculty Member at Carnegie Mellon University, Pittsburgh, PA, where he previously held the Strecker Early Career Professorship. His current broader research interests are in computer architecture, systems, hardware security, and bioinformatics. A variety of techniques he, along with his group and collaborators, has invented over the years have influenced industry and have been employed in commercial microprocessors and memory/storage systems. Mutlu has a BS in computer engineering and psychology from the University of Michigan, Ann Arbor, MI, and an MS and a PhD in electrical and computer engineering from the University of Texas at Austin, Austin, TX. He is a Fellow of the ACM and the IEEE and an Elected Member of the Academy of Europe (Academia Europaea).

■ Direct questions and comments about this article to Theocharis Theocharides, Department of Electrical and Computer Engineering, University of Cyprus, Nicosia 1678, Cyprus; ttheocharides@ucy.ac.cy.