Guest Editors' Introduction: Embedded Intelligence in the Internet-of-Things

Robert P. Dick
University of Michigan, Ann Arbor

Li Shang
University of Colorado Boulder

Marilyn Wolf
University of Nebraska–Lincoln

Shao-Wen Yang
Amazon.com, Inc.

THIS SPECIAL ISSUE focuses on the need for, and ways of achieving, embedded intelligence in the Internet-of-Things. After explaining why edge device and in-network intelligence are necessary, this editorial gives examples of relevant advances in efficient machine learning, wireless communication technology, and circuits. The survey is followed by four articles describing new results in this area.

In "Efficient Associative Search in Brain-Inspired Hyperdimensional Computing," Imani et al. report their findings in the area of biomimetic hyperdimensional computing, in which input vectors are projected into high-dimensional space to enable more efficient classification on resource-constrained systems. A direct comparison of the hypervectors associated with each sample and all of the possible class hypervectors would be computationally expensive. Therefore, the authors impose a hierarchy in which hypervectors are first assigned to categories. This enables the elimination of many vector compersions. The authors then describe an unusual quantization step to reduce the cost of each comparison. The valid values are tightly constrained, thereby limiting computational complexity, while still enabling low quantization error. In particular, they quantize the values of the following form: $2^i + 2^j$ for i and j in $\mathbb{Z}$. This results in a ten times improvement to energy efficiency and speed, compared with a baseline implementation of the hyperdimensional computing algorithm.

In "A Quantitative Exploration of Collaborative Pruning and Approximation Toward Energy-Efficient Deep Neural Networks," He et al. study the optimal tradeoff between numeric and network-level precisions in deep neural networks to optimize energy efficiency. They describe a two-stage algorithm. The first stage prunes layers with the most impact on energy consumption and the greatest accuracy resilience to pruning. The second stage uses stochastic gradient descent to guide the reduction of weight precisions to reduce energy consumption while preserving accuracy. Their pruning and weight quantization approach improves energy efficiency by 26% relative to weight optimization and 30% relative to pruning while improving accuracy.

In "Fune: An FPGA Tuning Framework for CNN Acceleration," Xiao and Liang describe a framework to optimize the use of dynamic reconfiguration of field-programmable gate arrays (FPGAs) to implement convolutional neural networks (CNNs). They first describe several design decisions influencing efficiency, including how well functionality can be batched to eliminate redundancy among configurations. They then decompose the problem into two subproblems: identifying the best positions in the network layer structure for reconfiguration, which they solve using a dynamic programming algorithm, and determining configurations enabling minimal latency within resource bounds, for which they use

Digital Object Identifier 10.1109/MDAT.2019.2957370
Date of current version: 6 February 2020.

a layer-by-layer search process. They evaluate their framework on several widely used CNNs, and report speedups ranging from 1.4× to 2.2×.

Finally, in "Tanji: A General-Purpose Neural Network Accelerator with a Unified Crossbar Architecture," Zhu et al. present a neural network accelerator appropriate for both convolutional and fully connected layers. Their system can be reconfigured to trade off numeric precision and parallelism by ganging pairs of 8-bit adder and comparator units into 16-bit units. Dataflow is carefully arranged to eliminate redundant actions including needless weight fetches, input feature transfers, and zero-weight operations. When prototyped on an FPGA, the system is competitive with several recent neural network accelerators implemented with application-specific integrated circuits. ■

■ Direct questions and comments about this article to Robert P. Dick, University of Michigan, Ann Arbor, MI, USA; dickrp@umich.edu.