Co-design for distributed AI. While AI technologies and applications have been advancing rapidly, AI algorithms are evolving from a centralized manner (cloud AI) to a distributed manner (edge AI), where the algorithms and distributed accelerators will work collaboratively. Such a new paradigm can largely alleviate the massive data storage and computation burden. The emerging privacy and security requirement of the AI solutions also brings in new design challenges. To adapt to such a paradigm shift, novel co-design methodologies are required for distributed AI development and deployment.

Co-design for heterogeneous and large scale AI. Another trend of future AI is heterogeneous and large scale. For example in smart city, the heterogeneity resides in all aspects including data, algorithms and devices, such as traffic planning, crowd monitoring, public health care, security, economy, and urban planning. In addition to heterogeneity, such applications may also involve millions or billions of edge nodes at an extremely large scale. To this end, novel co-design methodologies are essential to address these challenges and push AI to a new phase of real-world applications.

Co-design for emerging AI technologies. Both AI algorithms and devices are emerging dramatically. For example, the rapid development of process-in-memory technologies and the neuromorphic computing technologies have brought in immense opportunities. Moreover, the recent achievements of quantum computing have also opened a new era of AI development. Towards these interesting yet challenging new directions of future AI, revolutionary new design methodologies for software and hardware are mandatory.

ACKNOWLEDGMENT

This work is supported in part by the IBM-Illinois Center for Cognitive Computing System Research (C3SR) – a research collaboration as part of IBM AI Horizons Network, and Campus for Research Excellence and Technological Enterprise (CREATE) programme in Singapore.

REFERENCES

[1] Han Cai et al. Proxylessnas: Direct neural architecture search on target task and hardware. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
[2] Mingxing Tan et al. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[3] Bichen Wu et al. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
[4] Dustin Franklin. NVIDIA Jetson AGX Xavier delivers 32 teraops for new era of AI in robotics. *NVIDIA Accelerated Computing| Parallel For all*, 2018.
[5] Norman P Jouppi et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of International Symposium on Computer Architecture (ISCA)*, 2017.
[6] Yu-Hsin Chen et al. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. In *IEEE International Solid-State Circuits Conference (ISSCC)*, 2016.
[7] Xiaofan Zhang et al. High-performance video content recognition with long-term recurrent convolutional network for FPGA. In *Proceedings of the International Conference on Field Programmable Logic and Applications (FPL)*, 2017.
[8] Qin Li et al. Implementing neural machine translation with bi-directional gru and attention mechanism on FPGAs using HLS. In *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2019.
[9] Song Han et al. ESE: Efficient speech recognition engine with sparse lstm on FPGA. In *Proceedings of the International Symposium on Field-Programmable Gate Arrays (FPGA)*, 2017.
[10] Chuanhao Zhuge et al. Face recognition with hybrid efficient convolution algorithms on FPGAs. In *Proceedings of the Great Lakes Symposium on VLSI (GLSVLSI)*, 2018.
[11] Junsong Wang et al. Design flow of accelerating hybrid extremely low bit-width neural network in embedded FPGA. In *Proceedings of the International Conference on Field Programmable Logic and Applications (FPL)*, 2018.
[12] Xiaofan Zhang et al. DNNBuilder: an automated tool for building high-performance DNN hardware accelerators for FPGAs. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2018.
[13] Hanchen Ye et al. HybridDNN: A framework for high-performance hybrid dnn accelerator design and implementation. In *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, 2020.
[14] Cong Hao and Deming Chen. Deep neural network model and fpga accelerator co-design: Opportunities and challenges. In *Proceedings of the IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT)*, 2018.
[15] Cong Hao et al. NAIS: Neural architecture and implementation search and its applications in autonomous driving. *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2019.
[16] Cong Hao et al. FPGA/DNN co-design: An efficient design methodology for 1ot intelligence on the edge. In *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, 2019.
[17] Weiwen Jiang et al. Accuracy vs. efficiency: Achieving both through fpga-implementation aware neural architecture search. In *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, 2019.
[18] Yuhong Li et al. EDD: Efficient differentiable dnn architecture and implementation co-search for embedded AI solutions. *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, 2020.
[19] Xiaofan Zhang et al. SkyNet: a hardware-efficient method for object detection and tracking on embedded systems. In *Proceedings of Machine Learning and Systems (MLSys)*, 2020.
[20] Lei Yang et al. Co-exploration of neural architectures and heterogeneous asic accelerator designs targeting multiple tasks. *Proceedings of the ACM/IEEE Design Automation Conference (DAC)*, 2020.
[21] Hanxiao Liu et al. Darts: Differentiable architecture search. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.
[22] Dimitrios Stamoulis et al. Single-path nas: Designing hardware-efficient convnets in less than 4 hours. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019.
[23] Hardik Sharma et al. Bit fusion: Bit-level dynamically composable architecture for accelerating deep neural networks. In *ISCA*. IEEE, 2018.
[24] Sayeh Sharify et al. Loom: Exploiting weight and activation precisions to accelerate convolutional neural networks. In *DAC*. IEEE, 2018.
[25] Xiaowei Xu et al. DAC-SDC low power object detection challenge for UAV applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[26] Kaan Kara, Ce Zhang, and Gustavo Alonso. DAC-SDC'18 2nd place winner in FPGA track. https://github.com/fpgasystems/spooNN, 2018. Accessed: 2020-02-28.
[27] Kaan Kara and Gustavo Alonso. DAC-SDC'19 3rd place winner in FPGA track, 2019.
[28] Feng Xiong et al. DAC-SDC'19 2nd place winner in GPU track, 2019.
[29] Jianing Deng et al. DAC-SDC'19 3rd place winner in GPU track, 2019.
[30] Hao Lu et al. DAC-SDC'18 1st place winner in GPU track. https://github.com/lvhao7896/DAC2018, 2018. Accessed: 2020-02-28.
[31] Chuanqi Zang et al. DAC-SDC'18 3rd place winner in GPU track. https://github.com/xiaoyuuuuu/dac-hdc-2018-object-detection-in-Jetson-TX2, 2018. Accessed: 2020-02-28.
[32] Boran Zhao et al. DAC-SDC'19 2nd place winner in FPGA track, 2019.
[33] Shulin Zeng et al. DAC-SDC'18 1st place winner in FPGA track. https://github.com/hirayaku/DAC2018-TGIIF, 2018. Accessed: 2020-02-28.
[34] Cong Hao et al. DAC-SDC'18 3rd place winner in FPGA track. https://github.com/onioncc/iSmartDNN, 2018. Accessed: 2020-02-28.
[35] Xilinx. ChaiDNN. https://github.com/Xilinx/CHaiDNN.
[36] Mark Sandler et al. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
[37] Ningning Ma et al. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
[38] Lianghua Huang et al. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[39] Bo Li et al. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2019.
[40] Qiang Wang et al. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.