

Figure: Illustration of co-exploring neural architectures and hardware designs

As the increasing number of applications (ranging from mobile and connected home to security, surveillance and automotive) that combine edge computing and machine learning, implementing neural networks on edge devices become ever more important; however, with limited hardware resources on the edge devices and the restricted timing specification, neural networks that attain state-of-the-art accuracy will not fit. We establish that in order to achieve the highest accuracy and hardware efficiency on the edge, it is best to co-explore hardware design and neural architecture (the above figure), rather than taking state-of-the-art approaches that either explore neural architecture only assuming fixed hardware, or explore hardware design only assuming fixed neural architecture. The proposed co-exploration framework opens up the hardware design freedom, which can significantly push forward the Pareto frontiers between hardware efficiency and test accuracy, and we can obtain much better design trade-offs than the existing approaches. Our contribution can be seen as an analogy to the concept of hardware-software co-design, which has prevailed due to its superior.

We aim to develop a full-stack tools to support co-exploration of neural architecture and hardware design from (1) optimization of fixed neural network on FPGAs; (2) FPGA-implementation aware NAS; (3) co-explore neural architectures on multiple FPGAs; (4) co-explore neural architecture on network-on-chip-based scalable hardware.

[1] Weiwen Jiang, Lei Yang, Edwin Sha, Qingfeng Zhuge, Shouzhen Gu, Yiyu Shi and Jingtong Hu, "Hardware/Software Co-Exploration of Neural Architectures," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (under review).

[2] Weiwen Jiang, Xinyi Zhang, Edwin H.-M. Sha, Qingfeng Zhuge, Lei Yang, Yiyu Shi and Jingtong Hu, "Accuracy vs. Efficiency: Achieving Both through FPGA-Implementation Aware Neural Architecture Search," in Proc. of IEEE/ACM Design Automation Conference, 2019 (Nominated for Best Paper Award) (acceptance rate 25%)

[3] Weiwen Jiang, Edwin Sha, Xinyi Zhang, Lei Yang, Qingfeng Zhuge, Yiyu Shi and Jingtong Hu, "Achieving Super-Linear Speedup across Multi-FPGA for Real-Time DNN Inference," IEEE/ACM International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), New York, NY, 2019 (Nominated for Best Paper Award) (acceptance rate 29%).

[4] Lei Yang, Weiwen Jiang, Weichen Liu, Edwin Sha, Yiyu Shi, and Jingtong Hu, “Co-Exploring Neural Architecture and Network-on-Chip Design for Real-Time Artificial Intelligence”, in Proc. of the Asia and South Pacific Design Automation Conference, ASP-DAC, 2019.

[5] Qing Lu, Weiwen Jiang, Jingtong Hu and Yiyu Shi, "On Neural Architecture Search for Resource-Constrained Hardware Platforms," in Proc. of IEEE/ACM International Conference On Computer-Aided Design, Westminster, CO, Nov. 2019. (Invited Paper)

[6] Xinyi Zhang, Weiwen Jiang, Yiyu Shi and Jingtong Hu, "When Neural Architecture Search Meets Hardware Implementation: from Hardware Awareness to Co-Design," in Proc. of IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Miami, FL, Aug. 2019. (Invited Paper)