2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) | 978-1-6654-3555-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/FCCM51124.2021.00051

Heterogeneous Dual-Core Overlay Processor for Light-Weight CNNs

Tiandong Zhao, Yunxuan Yu, Kun Wang, Lei He
Electrical and Computer Engineering Department
University of California, Los Angeles

Convolutional neural networks (CNNs) have achieved extensive success on miscellaneous artificial intelligence applications such as image classification and object detection. A plethora of models emerge with different operators and architectures, gradually shifting attention from accuracy to efficiency in terms of speed and power, since VGG-like architecture from early stage has significant redundancy. Light-weight CNNs are proposed to reduce computation complexity and parameter amount. MobileNets, one typical example of light-weight CNNs, adopt depthwise separable convolution, while others such as SqueezeNet alter model topology to spare computation power.

However, such changes on operators and model topology bring about challenges to hardware efficiency. To be more specific, layers in modern light-weight CNNs are heterogeneous with regard to the computation pattern, especially between depthwise convolution layers and regular convolution layers. Even for the same layer type, layers show huge computation-and-communication (CTC) ratio differences that result from various layer characteristic parameters, e.g., input feature map size and kernel size.

There exist two paradigms in prior works on how to handle the heterogeneous CNN workload. One resorts to uniform architecture, while the other applies custom architecture to different models. Both show potential runtime PE inefficiency on the extra heterogeneity from depthwise convolutions. For the first paradigm, [1] adopts homogeneous processing elements (PEs) for different layer types, while [2] uses separate PEs. Yet, they do not provide or report any algorithm to automatically determine the PE sizes and PE numbers of their design. Since both allocate fixed number and size of PEs for various workloads, it will result in low runtime PE efficiency for different networks with different ratios between the regular convolution and the depthwise convolution. For the second paradigm, [3][4] accommodate various layer characteristics with tunable PE size, but no results on depthwise convolution are reported.

We propose a heterogeneous dual-core architecture (dual-OPU) and an algorithm searching for the best PE configuration to achieve high runtime PE efficiency for light-weight CNNs. For the architecture of the dual-OPU, one core is optimized for regular convolution layers and the other for depthwise convolution layers. PEs are homogeneous with each core. The PE configuration space includes the PE number for a core, the input size for each PE, and the line buffer, which realizes data reuse patterns for depthwise convolutions. For the searching algorithm, the search of the best PE configuration is guided by the number of multipliers allocated to two cores. Given a workload of a single CNN or multiple CNNs, we adopt branch-and-bound to locate the best multiplier ratio between two cores. Then the PE number, the input size for each PE, and line buffer size are determined locally for each core by exhaustive search over limited choices of PE sizes. For each PE configuration during the search, we interleave layers from two input images to balance the parallel workload on two cores. The layer allocation is further tuned by splitting along the input height dimension and reallocating the split workload to the other core. We adopt the PE configuration with the highest throughput constrained to limited resources.

We evaluate the dual-OPU on MobileNet v1, MobileNet v2, and SqueezeNet with a cycle-accurate instruction level simulator. The latency model and resource model used for searching and simulation provide $<3\%$ estimation error compared to real board-level measurement. Compared with a single-core processor with the same area for a single CNN, heterogeneous dual-OPU on average improves runtime PE efficiency and throughput by 11% and 31%, respectively. For a workload of multiple CNNs, dual-OPU improves average throughput by 11% compared with the state-of-the-art processors scaled to the same area.

References

[1] Yunxuan Yu, Tiandong Zhao, Kun Wang, and Lei He. Light-OPU: An fpga-based overlay processor for lightweight convolutional neural networks. In *The 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 122–132, 2020.

[2] DPU for Convolutional Neural Network 2019.

[3] Xuechao Wei, Yun Liang, Xiuhong Li, Cody Hao Yu, Peng Zhang, and Jason Cong. Tgpa: tile-grained pipeline architecture for low latency cnn inference. In *Proceedings of the International Conference on Computer-Aided Design*, pages 1–8, 2018.

[4] Xiaofan Zhang, Hanchen Ye, Junsong Wang, Yonghua Lin, Jinjun Xiong, Wen-mei Hwu, and Deming Chen. Dnnexplorer: a framework for modeling and exploring a novel paradigm of fpga-based dnn accelerator. In *Proceedings of the 39th International Conference on Computer-Aided Design*, pages 1–9, 2020.

978-1-6654-3555-0/21/$31.00 ©2021 IEEE
DOI 10.1109/FCCM51124.2021.00051