Optimized Distribution of an Accelerated Convolutional Neural Network across Multiple FPGAs

Alaa Maarouf, Nour El Droubi, Raghid Morcel, Hazem Hajj, Mazen A. R. Saghir and Haitham Akkary
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
{aim20, ngd02, rhm20, hh63, mazen, ha95}@aub.edu.lb

I. Introduction

Convolutional Neural Networks (CNN) have achieved a resounding success especially in computer vision and collaborative filtering. The general trend in CNN architectures has been to build deeper networks with a substantial number of convolution filters and several large feature maps. As a result, most of the current CNN inference routines are highly compute-intensive and have significant storage requirements. Field Programmable Gate Arrays (FPGAs) are among the most popular choices for accelerating CNN inference workloads as they can perform complex and massively parallel jobs. Recently, notable efforts have been made to distribute CNN inference workloads across multiple FPGAs [1]. These strategies, however, do not take into account variations in computational complexity across different layers of a CNN resulting in sub-optimal performance gains. This work proposes an optimal distribution of CNN layers across different FPGA nodes while accounting for each layer's performance to achieve maximum overall throughput.

II. Proposed Methodology

Our work draws inspiration from the deeply pipelined FPGA cluster design described in [1], which distributes the feed-forward computational phases of a CNN model across multiple FPGAs connected in a ring network topology. Our approach differs from [1] in two aspects. First, we make no assumptions about the cluster's topology, which enables any pair of FPGAs to communicate concurrently. Second, on every FPGA node, we deploy minimalist computation engines designed and optimized for resource-constrained devices as described in [2].

The CNN consists of N layers that need to be mapped to a set of K FPGA nodes. The objective is to optimally map the different computational layers of a given CNN to multiple FPGA nodes to achieve maximum throughput while taking into consideration the FPGA-to-FPGA communication bandwidth. Alternatively, we aim to reduce the computational latency of the slowest stage in the pipeline to obtain a more balanced stage distribution. The problem is redefined as deciding on the mapping of the individual CNN layers to different FPGA nodes while maximizing the minimum throughput throttling the pipeline. Our formulation is linearized by introducing artificial variables and can thus be solved using existing Integer Linear Programming (ILP) solvers.

III. Evaluation and Results

Our formulation can generalize to any CNN network. However, we used AlexNet to evaluate our approach due to its widespread use in the literature. We simulate AlexNet's feed-forward stages on Xilinx VC709 nodes. For simplicity, we assume homogeneous FPGA boards and that inter-FPGA communication employs Xilinx Aurora 8b/10b transceivers with a bandwidth of 750 MB/s. Moreover, we use the Gurobi optimization library to solve the formulated ILP problem, and we compute the best layer distributions for different FPGA cluster sizes.

Our initial results show that increasing the number of FPGA nodes in the cluster improves the overall throughput for AlexNet. However, as the number of FPGAs increases beyond six, the throughput saturates indicating that further attempts at parallelizing the execution of AlexNet layers will show little to no improvement in performance. With six FPGA nodes, we achieve a throughput of 2839.15 GOPS compared to 825.6 GOPS in [1] (3.4x speedup over prior state-of-the-art).

IV. Conclusion

Our proposed optimization-based approach to distributing CNN layers across FPGA nodes allows for more balanced pipeline stages that can improve overall performance. This formulation can generalize to any CNN network and FPGA cluster size. We simulated our approach with six FPGA nodes and achieved a 3.4x speedup over the prior state-of-the-art.

References

[1] C. Zhang, D. Wu, J. Sun, G. Sun, G. Luo, and J. Cong, "Energy-efficient cnn implementation on a deeply pipelined fpga cluster," in *Proceedings of the 2016 International Symposium on Low Power Electronics and Design*. ACM, 2016, pp. 326–331.

[2] R. Morcel, H. Hajj, M. A. R. Saghir, H. Akkary, H. Artail, R. Khanna, and A. Keshavamurthy, "Feathernet: An accelerated convolutional neural network design for resource-constrained fpgas," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 12, no. 2, pp. 6:1–6:27, Mar. 2019. [Online]. Available: http://doi.acm.org/10.1145/3306202

2576-2621/20/$31.00 ©2020 IEEE
DOI 10.1109/FCCM48280.2020.00068