Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M20-156P
Published: 3/18/2021

Inventors

Akshay Dua

Fengbo Ren

Contact

Shen Yan
shen.yan@skysonginnovations.com

A Scalable Parameterized OpenCL-Defined Accelerator Architecture for Efficient Convolutional Neural Network (CNN) Inference on FPGAs

Background

Convolutional neural networks (CNNs) are a class of deep neural network known for their superior extraction capability of shift/space invariant local features critical for high-level cognition tasks. CNNs are widely applied in image/video processing and computer vision tasks, including image/video recognition, object detection, and semantic segmentation, as well as medical image analysis and natural language processing. Due to their high computational intensity, CNN models require hardware acceleration for real-time applications. High-end graphics processing units (GPUs) are popular accelerators for CNNs. However, GPUs are not a preferred choice for energy- or thermal-constrained scenarios, such as Internet-of-things (IoT) and edge computing, as GPUs are power-hungry and have limited energy efficiency.

To enable real-time video analytics on a scale driven by IoT application growth, computations must happen at the network edge (near the cameras) in a distributed fashion. Recent studies have shown that field-programmable gate arrays (FPGAs) are highly suitable for edge computing due to the architecture's adaptiveness, high computational throughput for streaming processing, and high energy efficiency. The existing OpenCL-defined FPGA accelerators for CNN inference are insufficient due to limited flexibility for supporting multiple CNN models at run time and poor scalability resulting in underutilized FPGA resources and limited computational parallelism.

Invention Description

Researchers at Arizona State University have developed a scalable, parameterized, run-time-flexible OpenCL-defined accelerator architecture for efficient CNN inference on FPGAs. The CNN accelerator kernel features three user-defined architecture parameters that can be used to easily scale the computational parallelism, memory footprint, and memory bandwidth of the design according to a given FPGA system. This scalability allows for optimal usage of available hardware resources for any given FPGA system. The CNN accelerator kernel is also run-time-flexible in the context of multi-tenancy cloud/edge computing, which can be time-shared to accelerate a variety of CNN models at run time without the need of

recompiling the FPGA kernel hardware nor reprogramming the FPGA. Versatility is preserved by the CNN accelerator's generic form, which supports a wide variety of CNN models for numerous computer vision tasks, such as Alexnet, Retinanet, Resnet-50 for image classification, as well as Light-weight Retinanet for objective detection and video analytics.

Potential Applications

- Convolutional neural networks (CNNs)
- Computer vision
- Edge computing
- Internet-of-Things applications

Benefits and Advantages

- Highly parallelized and pipelined 1-D systolic array architecture delivers fast and energy-efficient CNN inference on FPGAs
- Parameterization of the design allows for the automated scaling of the design
- Easily adaptable to multiple CNN models at run-time without regenerating FPGA hardware
- Enables maximum utilization of the computing and memory resources of a given FPGA system

Research Homepage of Professor Fengbo Ren