Two Partially Reconfigurable Architectures for Efficient Implementation of Convolutional Neural Networks (CNNs)

Abstract

Today, convolutional neural networks (CNN) are widely used in many applications of artificial intelligence (AI), including image processing, video processing, natural language processing, and forecasting time series. CNNs require heavy computations to provide significant accuracy for many AI tasks. Normally, the training phase of CNNs is done through GPUs. However, the inference phase, which runs on the edge devices, is executed through a dedicated hardware accelerator. Therefore, the efficient implementations of CNNs to improve performance using limited resources without accuracy reduction is a challenge for AI systems for edge devices. One of the architectures for the efficient execution of CNNs is the array-based accelerator that consists of an array of similar Processing Elements (PEs). The array accelerators are popular as high-performance architecture using the features of parallel computing and data reuse. These accelerators are optimized for a set of CNN layers, not for individual layers. Using the same accelerator dimension size to compute all CNN layers with varying shapes and sizes leads to the resource underutilization problem. To handle this challenge, two research teams under my supervision propose two solutions namely CNNX and RASHT in last few years. In my talk, I will discuss both techniques and explain how we can have a flexible and scalable architecture for array-based accelerator that increases resource utilization. The RASHT architecture resizes PEs depending on the size and shapes of the layers. The CNNX implements a small and efficient accelerator and then using retiling executes the layers with different sizes and shapes. Both architectures have been evaluated through different CNN structures such as GoogLeNet, MobileNet and AlexNet. Experimental results show the effectiveness of both proposed hardware accelerators.