Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks

C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao and J. Cong

Summary of the Research Paper

Convolutional Neural Networks (CNN) - an extension of Artificial Neural Networks (ANN) - are a deep learning architecture with a plethora of applications, particularly in relation to image recognition. They work by loosely emulating the behaviour of optic nerves and brain synapses in biological creatures, processing data with multiple layers of neuron connections, to accurately recognise a given image.

Where general purpose processors lack efficiency in CNN implementation, Field-Programmable Gate Array (FPGA) based accelerators are an excellent enhancement having multiple advantages including high performance, high energy efficiency and even capability of reconfiguration.

However for any given CNN algorithm implementation, there are an abundance of potential solutions resulting in a vast design space. Given the constraints of computation resources and memory bandwidth, an accelerator structure that is not well constructed will have lower performance because of under-utilisation of logic resources or memory bandwidth.

The solution proposed by this paper is a roofline-model-based method for convolutional neural network's FPGA acceleration. This is done through means which will be elaborated on below:

- Optimising CNN's computational and memory access.
- Modelling all possible designs using a roofline model and discerning the best design for each layer.
- Finding the best cross layer design.
- Creating an actual implementation on a Xilinx VC707 board.

Key Points of the Research Paper

1) Exploring Multi-Layer CNN Accelerator Designs
    a) Achieving computation optimisation through methods of *loop unrolling*, *loop pipelining* and *tile size selection*.
    b) Achieving memory access optimisation through methods of *local memory promotion*, *loop transformations for data reuse* and *CTC radio*.
2) Implementation Details of the Solution
    a) Includes details of the proposed computer organisation (architecture) through *computation engines*, *memory sub-system* and *external data transfer engines*.
3) Experimental Results and Evaluation
    a) Setup of the experiments, and the experimental results that follow (regarding power & performance). Implemented with Vivado HLS on the VC707 board in C.

Proposed Questions:

➔ What other CNN application accelerator implementations are there? Is there a diverse range?
➔ Why focus on optimizing computation engines?
➔ Why is the roofline model a good approach to take?