Optimizing Memory Access with Shift-based On-Chip Memory for Efficient Convolution Processing in analog DNNs

Junaid Muhammad, Syed Asmat Ali, Malik Summair Asghar, Saad Arslan, HyungWon Kim*

Department of Electronics Engineering, Chungbuk National University, Cheongju,28644,South Korea

Abstract

Convolutional neural networks (CNNs) have emerged as a powerful tool for visual data processing, particularly in tasks like object detection. However, CNNs often require substantial computational resources, making real-time implementation a challenge [1]. In this paper, we present a novel approach to address the computational requirements of convolutional operations in the context of the DNNs. Our approach leverages a digital controller that efficiently processes 416 x 416 images using a minimal 7 x 5 memory footprint. By employing only three shift operations—left, right, and up—we significantly reduce data movement and optimize the access of overlapping pixels from dynamic random-access memory (DRAM). Figure 1 demonstrates data storage in on-chip memory, enabling a 28.57% reduction in memory size and a 12.5% decrease in memory access as compare to previous work [2].

Introduction

We propose a minimal on-chip memory capable of providing feature maps (FMAPs) and filter values to our analog convolutional neural network. By utilizing this memory architecture and shifting the pixels, we can avoid accessing off-chip memory for repeated pixels. After sending the 3x3 feature map/filter to the analog component, we move the pixels in either the left/right or up direction to obtain values for the next stride of the feature map. This methodology involves shifting the feature maps while keeping the filter fixed.

Through testing, we have determined that a 7x5 memory configuration which is shown in figure 1 is the most optimal. It can handle eight strides of data before requiring data loading for the next eight strides. Rather than flushing the existing data, we reuse the same data for subsequent strides until all rows have been processed. This approach allows us to load fewer feature maps for the next strides, improving efficiency.

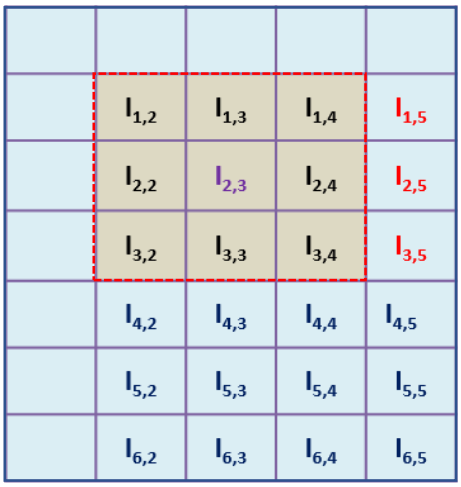

Fig 1. FMAP Storing pattern

Simulation Results

Figure 2 depicts our verification setup, which includes two interfaces for transmitting feature maps (FMAPs) and filters to the SoC. These interfaces allow us to utilize either the FPGA Processing system or Raspberry Pi to control our SoC. The simulation results, shown in Figure 3, demonstrate the operation of one accumulation unit on the analog side. Here, we send a 3x3 input image, along with a 3x3 filter and control signals from the digital controller, to perform a 3x3 convolution process followed by 2x2 pooling.

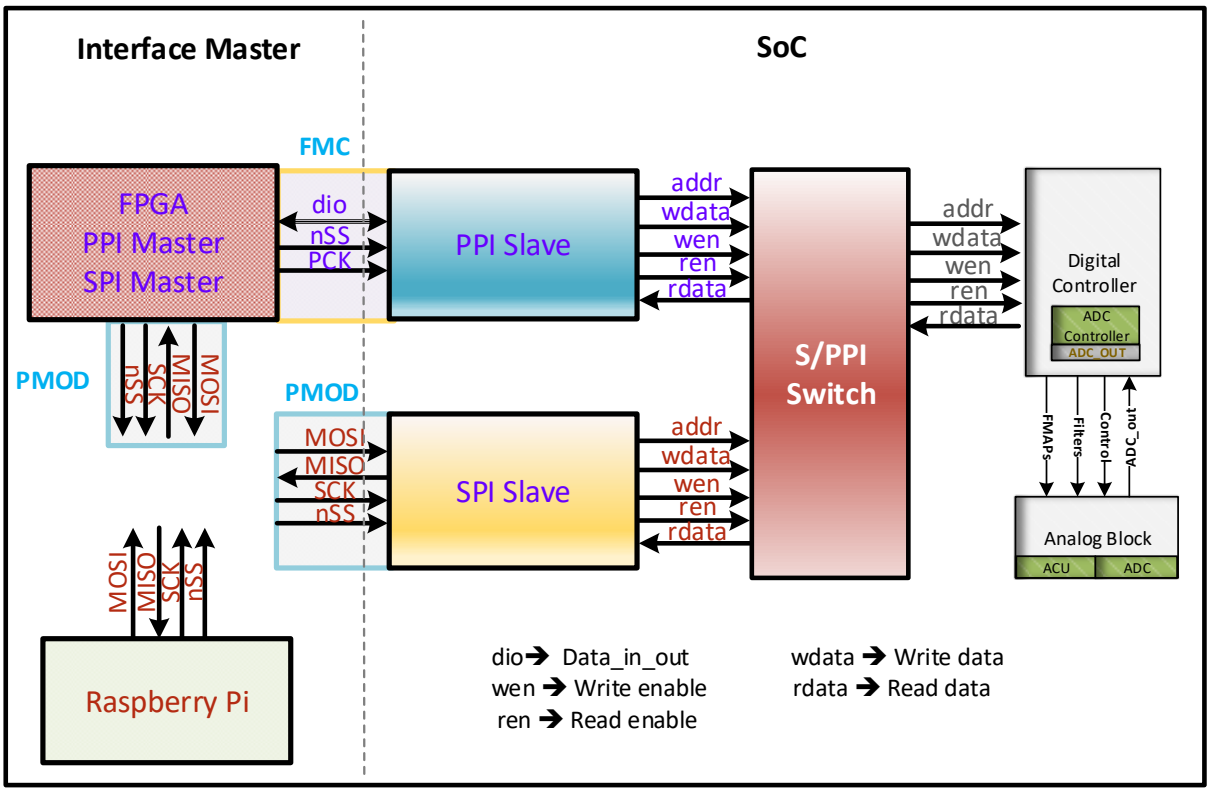

Fig 2. Verification Setup

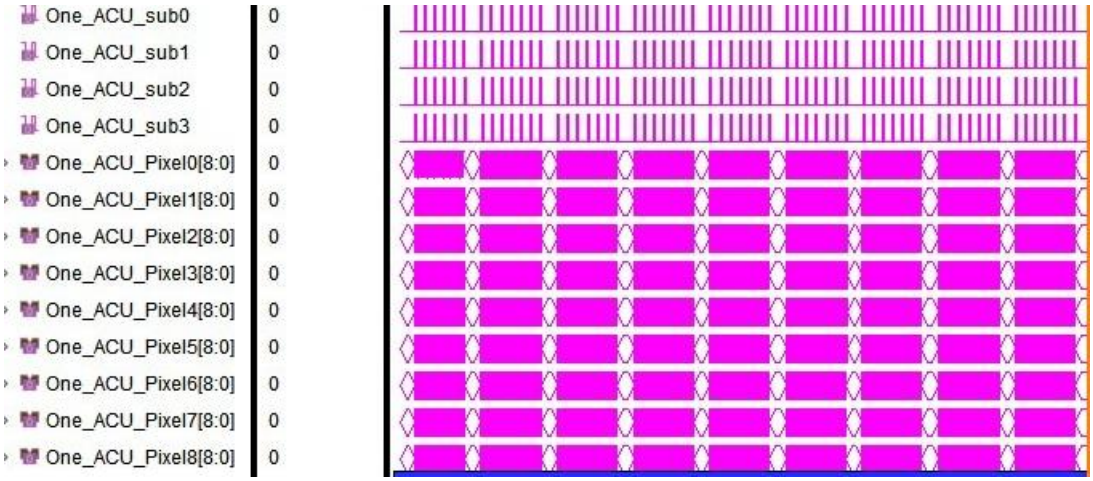

Fig 3. Simulation results for one ACU test setup

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant for RLRC funded by the Korea government (MSIT) (No. 2022R1A5A8026986, RLRC) and was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01304, Development of Self-learnable Mobile Recursive Neural Network Processor Technology). It was supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2023-2020-0-01462, Grand-ICT) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation). This research was supported by National R&D Program through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT(No. 2020M3H2A1076786, System Semiconductor specialist nurturing).

References

[1] Lee, Jungyeon & Asghar, Malik & Kim, HyungWon. (2023). A Low-Power 12-Bit SAR ADC for Analog Convolutional Kernel of Mixed-Signal CNN Accelerator. Computers, Materials & Continua. 75. 4357-4375. 10.32604/cmc.2023.031372.

[2] M. S. Asghar, M. Junaid, H. W. Kim, S. Arslan and S. A. Ali Shah, "A Digitally Controlled Analog kernel for Convolutional Neural Networks," 2021 18th International SoC Design Conference (ISOCC), Jeju Island, Korea, Republic of, 2021, pp. 242-243, doi: 10.1109/ISOCC53507.2021.9613851.