Live Demo: Memory-Efficient Hardware Design for a Real-Time Convolutional Encoder-Decoder Network

Min-Wu Jeong
Department of Electrical and Computer Engineering
Inha University
Incheon, Korea
22202243@inha.edu

Chan-Yong Shin
Department of Electrical and Computer Engineering
Inha University
Incheon, Korea
scyongg@inha.edu

Chae Eun Rhee
Department of Electrical and Computer Engineering
Inha University
Incheon, Korea
chae.rhee@inha.ac.kr

2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS) | 978-1-6654-0996-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/AICAS54282.2022.9870020

I. Introduction

This work presents a FPGA-based convolutional-neural-network (CNN)-based encoder-decoder accelerator for interpolation of high-resolution images. The baseline model is DVF [1]. The proposed system is demonstrated on Virtex UltraScale+ HBM VCU128 evaluation kit.

Fig. 1 (a) shows the overall architecture. It consists of three encoder layers, frame reuse module, three decoder layers and frame interpolation module, off-chip memory and memory manager performing data transfer between off-chip memory and accelerator. The frame reuse module avoids redundant computation between consecutive frames by recycling half feature map (fmap) channel. Frame interpolation module performs vector operation for frame synthesis using the optical flow data from the decoder. As shown in Fig. 1 (b), each layer consists of fmap loader, convolution buffer, buffer controller, fmap/weight fetcher, and processing element (PE) array. For layer pipelining, load and fetch of fmap must be performed simultaneously. Therefore, in Fig. 1 (c), the convolution buffer consists of double buffers. The output fmap of the previous layer is loaded into the convolution buffer of the following layer and fetched into the PE Array to perform multiplication and accumulation (MAC) operation.

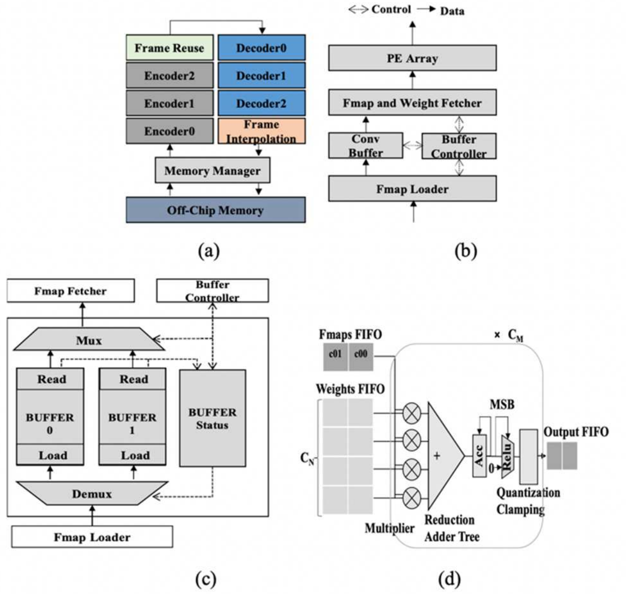

Fig. 1. The proposed encoder-decoder architecture with fused layers in row units (a) overall architecture (b) a layer architecture (c) convolution buffer (d) structure of processing element

Fig. 1 (d) shows the structure of the PE. It consists of a multiplier, a reduction adder tree, an accumulator, a ReLU, a quantization module and a clamping unit to cut a saturated value.

The frame interpolation architecture works as follows. The memory manager transfers input frames from off-chip memory to an accelerator. Optical flow is generated through the encoder-decoder layers. In-between frames are synthesized in a frame interpolation module by sampling pixel based on trilinear interpolation. The memory manager moves output results from accelerator to off-chip memory. Finally, the frame-rate up- converted video is displayed.

II. Demonstration

Fig. 2 shows FPGA setup environment for verification. The proposed system is demonstrated on Virtex UltraScale+ HBM VCU128 evaluation kit. 280K LUTs, 152K LUT RAM, 235K FF, and 5,280 DSP for multiplier are used. The performance of the proposed hardware is 1.4 TOPS with the operating clock frequency of 200MHz at 75% PE utilization. (42 GOPS×31) =1331 GOPS required for interpolating 2K@30fps videos to 60fps is sufficiently satisfied. The proposed system, which is optimized to be hardware-friendly, uses only about 25% of the parameters compared to the DVF model. The computational complexity was reduced to 17% thanks to the quantization considering the characteristics of each layer and the structure to avoid redundant calculations in processing consecutive frames in video. The PSNR degradation compared to DVF is less than 1 dB.

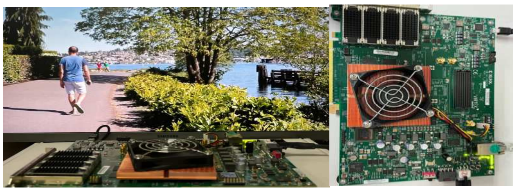

Fig. 2. FPGA verification environment

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2021R1A2C2008946).

References

[1] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, “Video frame synthesis using deep voxel flow,” in 2017 IEEE International Conference on Computer Vision (ICCV), 2017.