Parameter Embedding for FPGA Implementation of Binarized Neural Networks

Reina Sugimoto Nagisa Ishiura

School of Science and Technology, Kwansei Gakuin University

1 Introduction

A binarized neural network (BNN) is a restricted type of neural network where weights and activations are binary [1], which enables compact hardware implementation. Although many efficient architectures have been proposed [2, 3], they assume the weights and biases are stored in on-chip RAMs. This paper presents an attempt to embed those parameters into processing elements by utilizing LUTs in FPGAs as ROMs.

2 Binarized Neural Network (BNN)

For a neuron i in a neural network, let I_i be a set of neurons feeding i, b_i be the bias of i, $w_{i,j}$ be the weight associated with the synapse connecting neurons j and i. The activation value a_i of i is expressed as $a_i = f(b_i + \sum_{j \in I_i} w_{i,j} \cdot a_j)$. In the binarized neural network, a_i and $w_{i,j}$ are in $\{-1, +1\}$, and $f(x) = 1$ if $0 \leq x$ and $f(x) = -1$ otherwise. If we represent -1 and $+1$ as 0 and 1, respectively, the multiplication in the above formula is reduced to an exclusive-nor operation. To accommodate a huge neural network in a single chip, the weight and bias parameters are usually placed in on-chip RAMs.

3 Parameter Embedding

Instead of storing the weight and bias parameters in RAMs, we embed them in processing elements (PEs) for neurons, which increases freedom of overall hardware architecture. In this article we present an architecture for the two dimensional convolution layer.

Fig. 1 shows the structure of the PE for a neuron, which accumulates the products of the weights and the input activations of synapses sequentially. The weight parameters are stored in an unit w. When implementing a 25-input neuron on an FPGA, one LUT of 5 inputs is enough. The address a to choose the weights is supplied as an input. Comparison is done against a constant threshold c. Fig. 2 is the overall architecture for the convolution of the first layer of LeNet [4]. It is a 32×32 two dimensional array with rotate shift capability where 28×28 of them are PEs. A feature map (32×32 pixels) is loaded to the array in the first 32 clocks from the upper side, then all the PEs do computation for their first synapse in parallel. Next, the map is (rotate) shifted left by 1 bit to feed the second inputs to PEs. This is repeated 4 times. After that the map is (rotate) shifted upwards. Then the map is shifted right 4 times, shifted upwards, shifted left 4 times, shifted upwards, and so on, so that all the 25 inputs are accessed. This takes 25 cycles. After writing back the result into x, the map data are shifted out downwards.

4 Preliminary Result

A module in Fig. 2 has been designed in Verilog HDL, where the weights and biases were determined randomly. The result of logic synthesis targeting Xilinx FPGA Artix-7 (xc7a100tcsg324-3) by Vivado (2016.4) is summarized in Table 1. The average LUT count per PE is about 9.38. This is because the counter in each PE is optimized according to constant parameter b.

5 Conclusion

We have presented a parameter embedded architecture for convolutional BNNs. We are now working on implementing a whole LeNet circuit on a FPGA chip.

References

[1] M. Courbariaux, et al.: “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1,” Computer Research Repository (Mar. 2016).

[2] K. Ando, et al.: “BRein memory: A single-chip binary/ternary reconfigurable in-memory deep neural network accelerator achieving 1.4 TOPS at 0.6 W,” *IEEE J. Solid-State Circuits*, vol. 53, no. 4, pp. 983–994 (Apr. 2018).

[3] H. Nakahara, et al.: “A memory-based realization of a binarized deep convolutional neural network,” in *Proc. FPT 2016*, pp. 273–276 (Dec. 2016).

[4] Y. LeCun, et al.: “Gradient-based learning applied to document recognition,” *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324 (Nov. 1998).

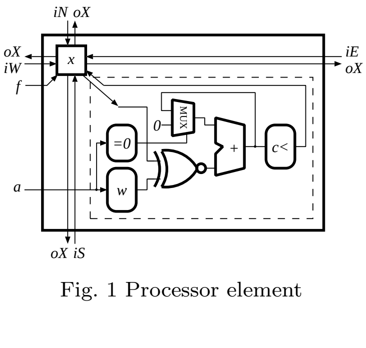

Fig. 1 Processor element

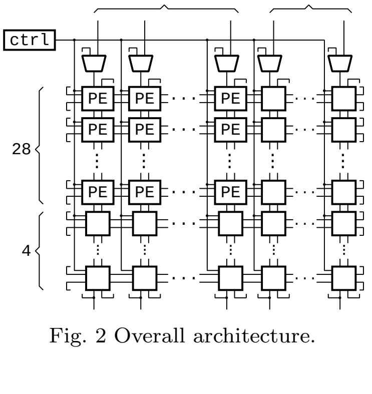

Fig. 2 Overall architecture.

Table 1 Synthesis result

	#FF	#LUT	delay [ns]
overall	5,003	7,591	3.419

Synthesizer: Xilinx Vivado (2016.4)
Target: Xilinx Artix-7