14.2 DNPU: An 8.1TOPS/W Reconfigurable CNN-RNN Processor for General-Purpose Deep Neural Networks

Dongjoo Shin, Jinmook Lee, Jinsu Lee, Hoi-Jun Yoo

KAIST, Daejeon, Korea

Recently, deep learning with convolutional neural networks (CNNs) and recurrent neural networks (RNNs) has become universal in all-around applications. CNNs are used to support vision recognition and processing, and RNNs are able to recognize time varying entities and to support generative models. Also, combining both CNNs and RNNs can recognize time varying visual entities, such as action and gesture, and to support image captioning [1]. However, the computational requirements in CNNs are quite different from those of RNNs. Fig. 14.2.1 shows a computation and weight-size analysis of convolution layers (CLs), fully-connected layers (FCLs) and RNN-LSTM layers (RLs). While CLs require a massive amount of computation with a relatively small number of filter weights, FCLs and RLs require a relatively small amount of computation with a huge number of filter weights. Therefore, when FCLs and RLs are accelerated with SoCs specialized for CLs, they suffer from high memory transaction costs, low PE utilization, and a mismatch of the computational patterns. Conversely, when CLs are accelerated with FCL- and RL-dedicated SoCs, they cannot exploit reusability and achieve required throughput. So far, works have considered acceleration of CLs, such as [2-4], or FCLs and RLs like [5]. However, there has been no work on a combined CNN-RNN processor. In addition, a highly reconfigurable CNN-RNN processor with high energy-efficiency is desirable to support general-purpose deep neural networks (DNNs).

In this paper, we present an 8.1TOPS/W reconfigurable CNN-RNN processor with the following 3 key features: 1) A reconfigurable heterogeneous architecture with a CL processor (CP) and a FC-RL processor (FRP) to support general-purpose DNNs, 2) a LUT-based reconfigurable multiplier optimized for the dynamic fixed point with on-line (on-chip) adaptation via overflow monitoring to exploit maximum efficiency from kernel reuse in the CP, 3) a quantization table (Q-table)-based matrix multiplication to reduce off-chip memory access and remove duplicated multiplications in the FRP.

The overall architecture of the proposed deep neural processing unit (DNPU) is shown in Fig. 14.2.2. It consists of the CP, FRP, and a top RISC controller. The CP is composed of 4 convolution clusters and 1 aggregation core. Each convolution cluster performs convolution operations with 4 convolution cores, and transfers the accumulation results to the accumulation core. One convolution core contains 4 PE groups with 48 PEs. The FRP performs matrix multiplication with the 128-enty Q-table, and 8 16b fixed-point multipliers are used to update the Q-table. The CP and FRP are able to process 4 different CLs and 8 RLs, respectively, in parallel.

The proposed input-layer division method (mixed division) is shown in Fig. 14.2.3. Due to limitations in on-chip memory size, the input layer image must be divided into several parts. There are three possible division methods: image division (ID), channel division (CD) and mixed division (MD). In the case of ID, the same weight parameters must be loaded multiple times for each divided image. For the CD, final output elements cannot be calculated with a single divided image, therefore multiple off-chip accesses are required. These two methods combine in the MD, which uses both ID and CD. In the MD, image division # and channel division # are selected to minimize the off-chip accesses. For various channel divisions, the proposed CP supports various channel #s (16 to 1024) and image sizes (32×16 to 256×128) with 5 different accumulation hierarchies (cluster, core, memory, bank and in-bank division). Supported configurations can be processed without any degradation of PE utilization.

Figure 14.2.4 shows the proposed layer-by-layer dynamic fixed-point with on-line adaptation architecture, and its optimized LUT-based reconfigurable multiplier. Previous work [3] took advantage of dynamic fixed-point to reduce the word length. In this work, further reductions can be achieved with the on-line adaptation via overflow monitoring. If the current output MSB cannot handle the final accumulation results from the 30b accumulation paths, the overflow count is increased. It is repeated for the 2nd MSB. Each overflow count is compared with the threshold (in this case 0.2%), and then the new fraction length is calculated to fit the desired overflow ratio. With this scheme, we can achieve 66.3% top-1 accuracy with 4b word lengths, while 32b floating point shows 69.9%. In the convolution operations, multiplication with the same weight is repeated 100-100,000 times. Leveraging this property, LUT-based multiplication can be constructed with an 8-entry physical LUT and 8-entry logical LUT. Also, four 4b multiplication and two 8b multiplication results are attained without additional cost, and those can be used for the respsective word lengths.

The detailed architecture of the Q-table-based FRP is shown in Fig. 14.2.5. The FRP has a reconfigurable 127-entry Q-table. It can function as one 7b Q-table to eight 4b Q-tables. Each entry of the Q-table contains the pre-computed multiplication result between a 16b fixed-point input and a 16b fixed point weight. After the Q-table is constructed once, only quantized indexes are required to compute the product. With the Q-tables, off-chip accesses can be reduced by 75%, and 99% of the 16b fixed-point multiplications can be avoided.

Measurement results are shown in Fig. 14.2.6. The DNPU can operate from 0.765-1.1V supply with 50-200MHz clock frequency. The power consumption at 0.765V and 1.1V are 34.6mW and 279mW, respectively. For a particular frame rate, energy-efficiency and bit-width (accuracy) can be traded off with one another. For the CP, word length can be changed from 4b to 16b, and quantization bit-width can be configured from 4b to 7b. With the proposed layer-by-layer on-line adaptation via overflow monitoring, the fraction length for each layer is adjusted rapidly without any off-chip learning for dynamic fixed-point. The softmax score of the answer object increases with the fraction length adaptation. If the scene is changed to another object, the adaptation flow is invoked for a new fraction length. As shown in the graph, 32b floating point precision is achieved with only 4b word length.

The DNPU shown in Fig. 14.2.7 is fabricated using 65nm CMOS technology and it occupies 16mm^2 die area. The proposed DNPU is the first CNN-RNN SoC with the highest energy efficiency (8.1TOPS/W). The table shows a performance comparison with the 3 previous deep learning SoCs. This work is the only one that supports both CNNs and RNNs. Compared to [2] and [3], this work shows 20× and 4.5× higher energy efficiency, respectively. Also, DNPU shows 6.5× higher energy efficiency compared to [5].

References:

[1] O. Vinyals, et al., "Show and Tell: A Neural Image Caption Generator," *Computer Vision and Pattern Recognition*, pp. 3156-3164, 2015.
[2] Y. Chen, et al., "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *ISSCC*, pp. 262-263, 2016.
[3] B. Moons, et al., "A 0.3-2.6TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets," *IEEE Symp. on VLSI Circuits*, 2016.
[4] J. Sim, et al., "A 1.42TOPS/W Deep Convolutional Neural Network Recognition Processor for Intelligent IoT Systems," *ISSCC*, pp. 264-265, 2016.
[5] S. Han, et al., "EIE: Efficient Inference Engine on Compressed Deep Neural Network," *IEEE/ACM Int'l Symp. on Computer Arch.*, pp. 243-254, 2016.

 978-1-5090-3758-2/17/$31.00

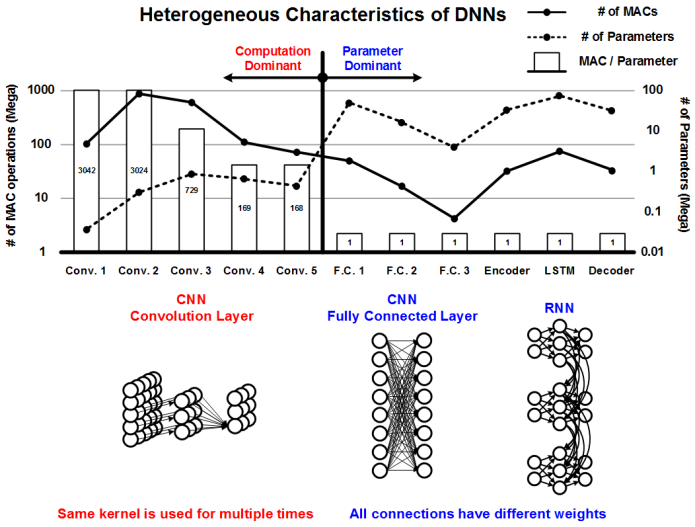


Figure 14.2.1: Heterogeneous characteristics of deep neural networks.

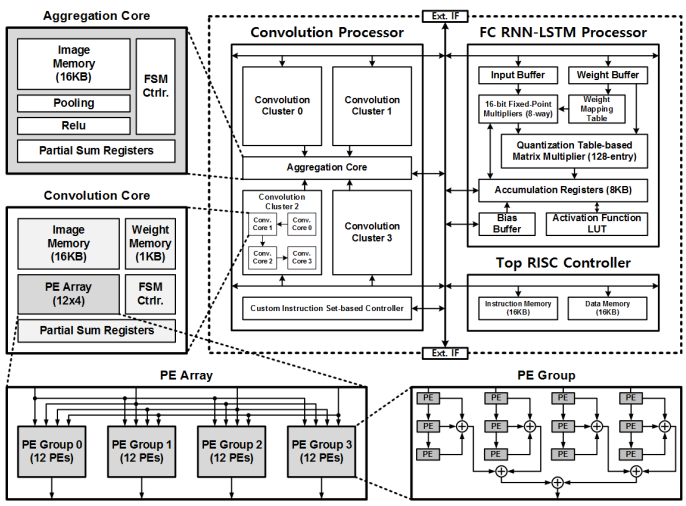


Figure 14.2.2: Overall architecture of the proposed CNN-RNN processor (DNPU).

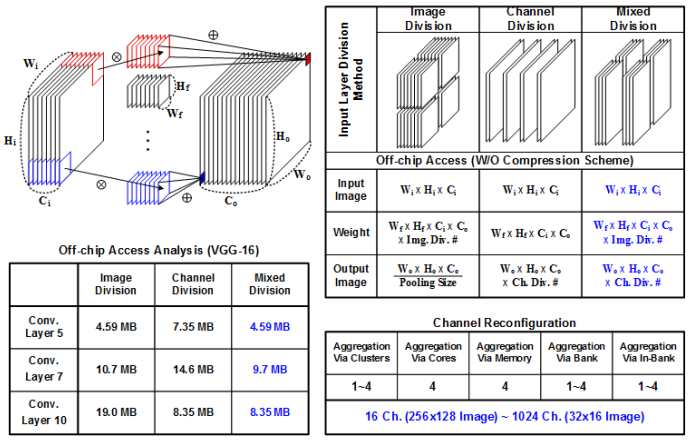


Figure 14.2.3: Mixed division methods and reconfigurability.

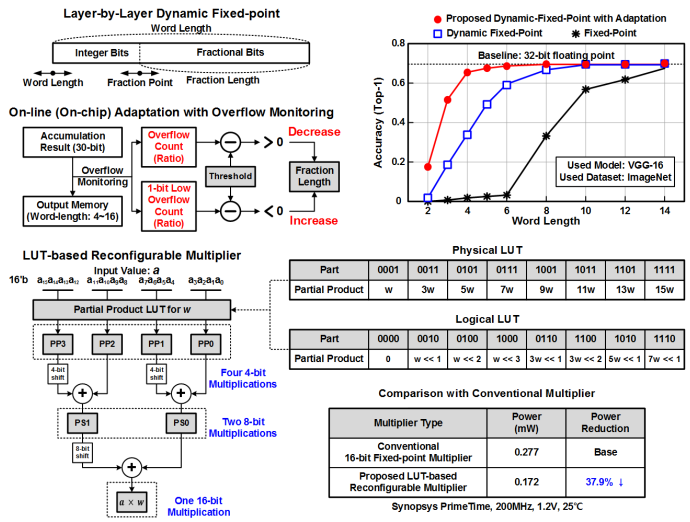


Figure 14.2.4: Layer-by-layer dynamic fixed-point with on-line adaptation and optimized LUT-based multiplier.

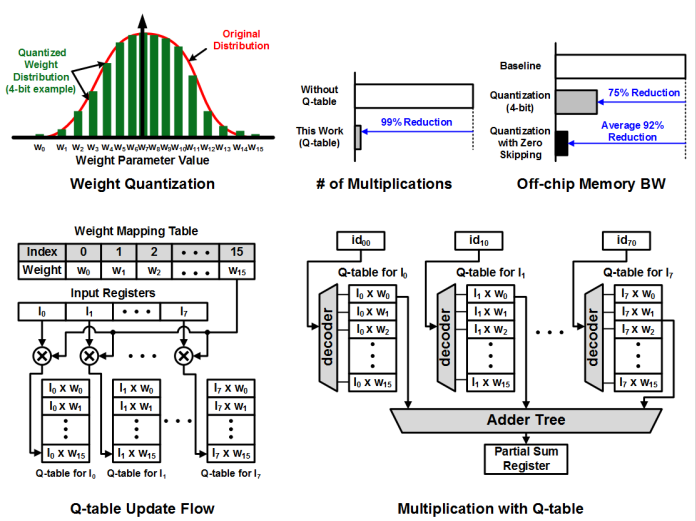


Figure 14.2.5: Quantization table-based matrix multiplication.

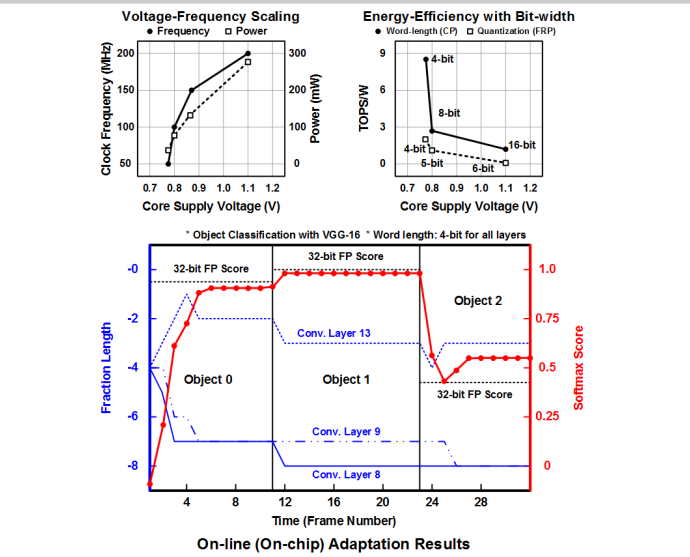
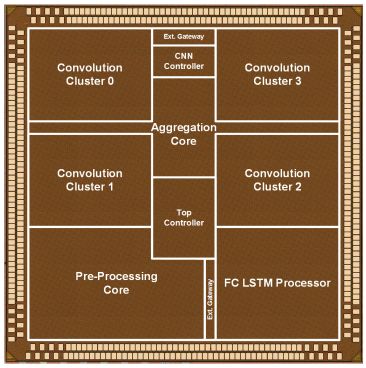


Figure 14.2.6: Measurement results.



	Specifications			
Technology	65nm 1P8M CMOS			
Die Area	4mm x 4mm (16mm^2)			
SRAM Size	CLP	FCRLP		
	280 KB	10 KB		
Supply Voltage	0.77V ~ 1.1V			
Operating Frequency	CLP	FCRLP	NoC	ETC
	~200MHz	~200MHz	~400MHz	~200MHz
Power Consumption		CLP	FCRLP	ETC	Total
	50MHz @ 0.77V	29mW	2.6mW	3mW	34.6mW
	200MHz @ 1.1V	235mW	21mW	23mW	279mW
Energy Efficiency		CLP Word Length : 16-bit		CLP Word Length : 4-bit	
	50MHz @ 0.77V	2.1 TOPS/W		8.1 TOPS/W	
	200MHz @ 1.1V	1.0 TOPS/W		3.9 TOPS/W	

	General					CLP						Quant. Bitwidth (bit)	FCRLP						
	Tech. (nm)	Clock Freq. (MHz)	Supply Voltage (V)	CNN Support	RNN Support	Word Length (bit)	Num. of MACs	Peak Perform. (GOPS)	AlexNet				Num. of MACs	Peak Perfor. (GOPS)	Power (mW)	Energy Efficiency (TOPS/W)	AlexNet	
									Power (mW)	Energy Efficiency	Frame Rate						Power (mW)	Frame Rate
ISSCC 2016 [2]	65	100 – 250	0.82 – 1.17	o	x	16	168	84	278 (200MHz)	0.21 TOPS/W	34.7 fps	-	-	-	-	-	-	-
SoVC 2016 [3]	40	12 – 204	0.55 – 1.1	o	x	1 – 16	256	102	76 (~100MHz)	0.94 TOPS/W	47 fps	-	-	-	-	-	-	-
ISCA 2016 [5]	45	800	100 – 250	x	o	-	-	-	-	-	-	4	64	100	590	0.17	600	18.8k fps
Proposed	65	50 – 200	0.77 – 1.1	o	o	4 – 16	768 (16b) 3072 (4b)	300 (16b) 1200 (4b)	63 (100MHz)	4.2 TOPS/W	177 fps	4 – 7	64	25	21	1.1	3.5	1.2k fps

Figure 14.2.7: Chip photograph and performance summary.

978-1-5090-3758-2/17/$31.00 ©2017 IEEE