A Novel Vision Chip Architecture for Image Recognition Based on Convolutional Neural Network

Honglong Li, Zhongxing Zhang, Jie Yang, Liyuan Liu, and Nanjian Wu*

State Key Laboratory for Superlattices and Microstructures, Institute of Semiconductors
Chinese Academy of Sciences, Beijing, 100083, China
* Email: nanjian@red.semi.ac.cn

Abstract

This paper presents a novel vision chip architecture for high accuracy image recognition based on the state-of-the-art algorithm – convolutional neural network (CNN). The architecture consists three hierarchical parallel processors: a processing element (PE) array, a row processor (RP) array and a dual-core microprocessor (MPU). It is compatible with conventional algorithms and reconfigurable for computing convolutional neural networks effectively. The architecture was implemented on a FPGA platform with 50MHz system clock, it achieves high classification accuracy up to 96.3% and high frame rate more than 1600fps. Experiment results indicate that the vision system can achieve real-time performance for image recognition applications.

1. Introduction

Computer vision has been applied in many fields, such as robot vision system, autonomous vehicle control and video surveillance. The object recognition is the most important task in these application. Usually, a vision based intelligent robot or an autonomous vehicle control system need to analyze scenes and recognize target objects they see in the image sensor, and then decide how to respond. So these applications often require recognition algorithms that have high degree of accuracy while being able to execute in real-time.

Vision chip [1, 2] is a device integrating image sensor and parallel image processor on a single chip. It can perform the traditional algorithms that include feature extraction and classification [3, 4]. The system overcomes serial image transmission and serial image processing bottlenecks and achieves high processing speed. But the traditional algorithms are shallow learning models and hard to achieve high accuracy. Recently, convolutional neural network (CNN) is widely used in image recognition because its high recognition accuracy [5, 6]. Such a deep network is computationally very expensive. Current vision systems are hard to run the network models in real-time with low power consumption [7].

In this paper, we propose a novel vision chip architecture and implement simple convolutional neural networks used for image recognition. The architecture consists of three hierarchical parallel processors: a processing element (PE) array, a row processing unit (RP) array and a two-core microprocessor (MPU). It can execute traditional image processing algorithm rapidly and reconfigure for computing convolutional neural networks effectively.

The rest of this paper is organized as follows: In section 2, the vision chip architecture are presented. In section 3, the convolutional neural network algorithm and Implementation is introduced. In section 4, the experiments and their results are described. Final, the conclusion is draw in section 5.

2. Architecture of Vision Chip

A block diagram of the proposed vision chip architecture is shown in Figure 1. The proposed vision chip has these main components: an image sensor or external data memory, hierarchical processor array, some necessary control logic circuits, clock generation circuits and buses. The hierarchical processor comprises three processing cores of different capabilities: an N×N pixel-level parallel PE array, an N×1 RP array and a dual-core MPU. The PE array is a two dimension computational array and can work in a single instruction multiple data (SIMD) way. Each PE consists of a 1-bit ALU with the computing power of arithmetic, logical and conditional operations in a single clock cycle. The operate data are from the current PE or the adjacent four PEs. The PE array mainly used for the low-level and image's local processing, such as 2D-filters, background reduction. The RP array is more processing power compared with the PE array. Each RP can complete a 16-bit logic operation. With this complex computational power, it easy complete histogram statistics, vector merge. The two-core MPU manages the whole system but also has the ability of serial algorithms.

The N×N PE array can be reconfigured into (N/4)×(N/4) sub-array when computing neural network. Every sub-array composed of 4×4 PEs is a neuron, see in Figure 2. The 16 PEs in one neuron is chained in a snake style, so that the neuron is reconfigured into a 16-bit processing unit. The reconfigured new neuron array is more convenient when processing multi-bit data.

978-1-4799-8485-5/15/$31.00 ©2015 IEEE

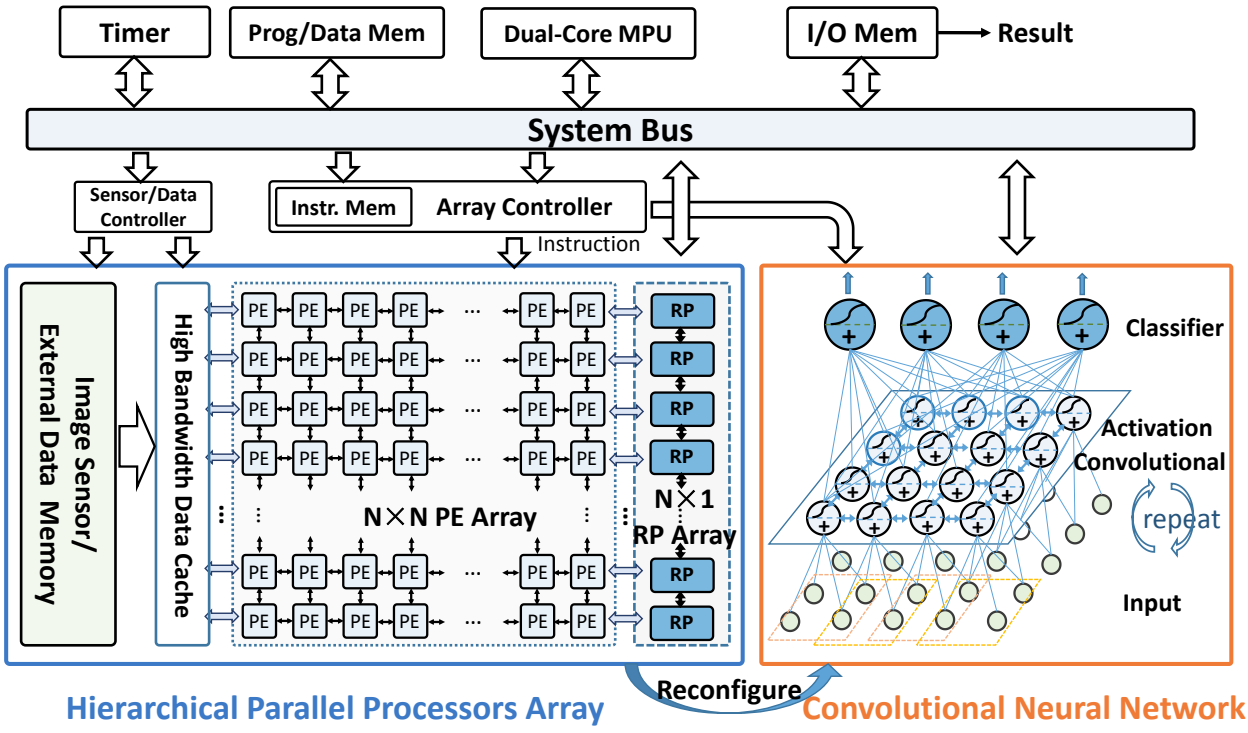

Figure 1. The vision chip architecture.

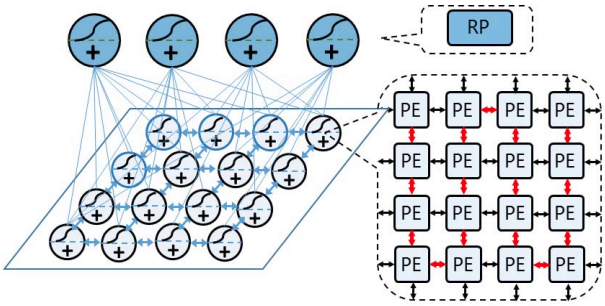

Figure 2. Neuron reconfiguring

3. Convolutional Neural Network Implementation

Convolutional neural network is composed of multiple layers which are inspired models of the human visual system. As with many other machine learning algorithms, the convolutional neural network model is trained based on a lot of training data. This training process called learning can be completed offline. The computation of convolutional neural network is huge but regular, especially suitable for visual chip that work in a single instruction multiple data (SIMD) way. Figure 3. shows the architecture of network.

3.1 Convolutional layer

Convolutional layer is using convolution kernels to extract features from upper layer and generate feature maps. Convolution is a neighboring operation, every result is only related to the several adjacent pixels. So it is easy to execute concurrently. In the interconnected PE array, every neuron calculates independently. We can get the convolution results quickly due to the high parallelism.

With the increase of network layers, the size of feature map exponentially narrow but the amount exponentially multiply (see in Figure 3). We divide the array into different particle size. When calculating the Cov1 layer, we need the whole PE array just like Figure 4(a). When calculating the Cov2 layer, feature maps become small. We divide PE array into four parts and calculate four maps concurrently, just like Figure 4(b). The operations in four parts are the same but input data and convolution coefficients are different. In next convolutional layer, the particle size will be smaller, divided into 16 parts, 64 pats. So that we can make full use of computing resources.

3.2 Activation layer

Activation function in neural networks is very common, it can optimize network performance. The most famous activation function may be the Sigmod Function shown in Figure 5(a). But it is a nonlinear function. Avoiding nonlinear computation and lookup table, we use a simple activation function shown in Figure 5(b). We just need to judge positive or negative which is easy for PE array.

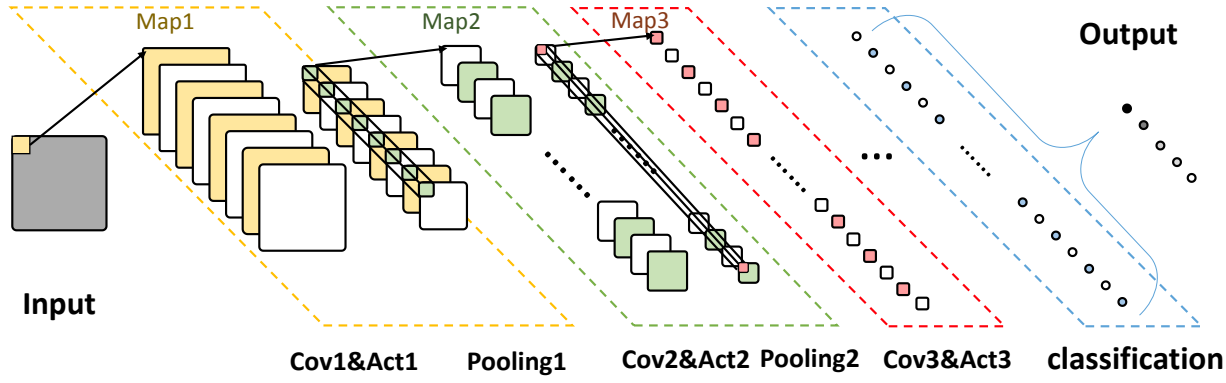

Figure 3. Convolutional neural network

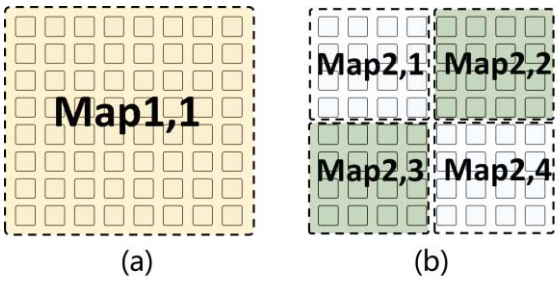

Figure 4. Particle size

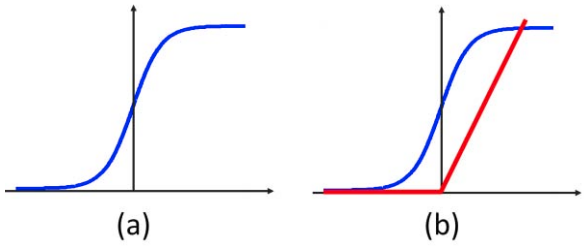

Figure 5. Activation function

3.3 Pooling layer

Pooling layer can reduce the amount of data and introduce translation or deformation invariance. The pooling procedure is shown in Figure 6. We divide the feature map into small cells overlap or not. Then find the max or average value in every cell. In down sampling step, we use a special instruction to skip some PEs when store feature map to memory.

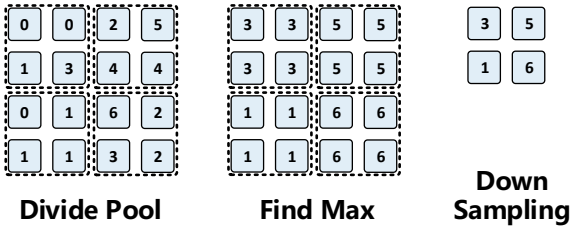

Figure 6. 2×2 max pooling

These three kinds of layers pile repeatedly and transform the image into a long feature vector. Finally we use a classification layer as output layer. Softmax classifier is a usually option in neural networks. Most of classifier is to calculate the inner product of feature vector and classifier parameters. The stronger RP array is suitable for classification layer.

So far, we implement the convolutional neural network on our vision chip architecture completely.

4. Experiment and performance

In this section we will give two experiments on handwritten number recognition and face detection.

Hand-writing number recognition is a multi-category task. The data for test were taken from the MNIST (Modified National Institute of Standards and Technology) dataset. There are 0~9 ten kinds of handwritten digits in the image dataset. We built a convolutional neural network contained three convolutional layers. The data type of the input image and feature maps is 8bit integer while the data type of convolution kernel is 8bit fix-point decimal. The network scale is shown in Table 1.

We use 60000 images as training data and other 10000

Table 1. Network scale in number recognition

Layer	Map Size	Map amount	Kernel Size
Cov1	28×28	4	5×5
Act1	24×24	4	--
Poo1	24×24	4	2×2
Cov2	12×12	16	5×5
Act2	8×8	16	--
Poo2	8×8	16	2×2
Cov3	4×4	64	4×4
Act3	1×1	64	--
Softmax	64×1	10	64×1

images as test data. Every image cost about 30000 instruction cycle. Under 50MHz system clock, it can processing 1666 images per second. The most instruction cycle is used on multiplication. If we integrate multiplier in PE array, the performance will increase significantly .The classification accuracy is 96.3%. Figure 7. shows some recognition result.

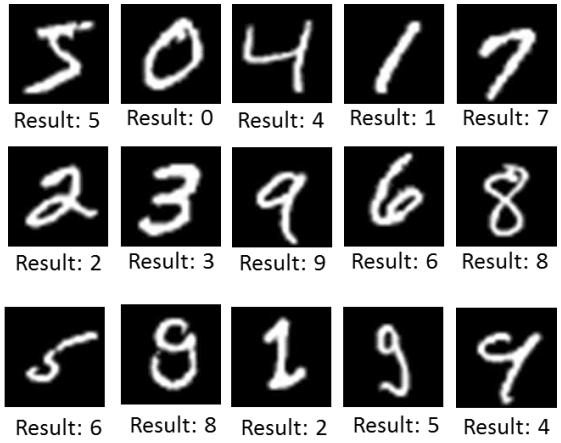

Figure 7. Number recognition

Face detection is a dichotomy task. It is simpler than number recognition because we just judge is or not. We built a simpler network that scale halve. Every 3232 image cost about 15000 instruction cycle. We can search in 256×256 image at speed of 52fps. The detection rate is 94.5%. Figure 8. shows some recognition result.

Figure 8. Face detection

The vision chip architecture is implemented using an Altera Arria V FPGA board. The system is running under 50MHz system clock. The PE array and the RP array respectively contains 128×128 PEs and 128×1 RPs. The processor array can be feed with real world images though an on-board image sensor or receive image from external I/O memory. The detail performance is shown in Table2

Table 2. Performance

Clock Freq.	*50MHz*
Image size	32×32
Conventional algorithms	LBP, HOG Adaboost
non-von Neumann neural networks	CNN
Detection accuracy	96.3%.
Performance	>1600 fps @32×32 >50 fps @256×256

5. Conclusion

In this paper, a novel vision chip architecture for enabling real-time execution of convolutional neural network was introduced. The architecture is implemented on FPGA platform under 50MHz frequency and achieves classification accuracy up to 96.3% and high frame rate more than 1600fps. Experiments indicate that the vision system can achieve real-time performance for image recognition applications.

References

[1] Miao W, Lin Q, Zhang W, et al. A programmable SIMD vision chip for real-time vision applications[J]. Solid-State Circuits, IEEE Journal of, 2008, 43(6): 1470-1479.

[2] W. C. Zhang, Q. Y. Fu, and N. J. Wu, "A programmable vision chip based on multiple levels of parallel processors," *IEEE J. Solid-State Circuits,* vol. 46, no. 9, pp. 2132-2147, Sept. 2011.

[3] C. Shi, J. Yang, N. J. Wu and Z. H. Wang "A 1000 fps Vision Chip Based on a Dynamically Reconfigurable Hybrid Architecture Comprising a PE Array Processor and Self-Organizing Map Neural Network", IEEE J. Solid-State Circuits, vol.49, no. 9, pp. 2067 - 2082 Volume: 49, Sept. 2014.

[4] J. Yang, C. Shi, and N. J. Wu, "Heterogeneous vision chip and LBP-based algorithm for high-speed tracking", Electronics Letters, vol. 50, no. 6, pp. 438-439, 2014.

[5] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 05 2015.

[6] van Doorn J. Analysis of Deep Convolutional Neural Network Architectures[J]. 2014.

[7] Zhang C, Li P, Sun G, et al. Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks[C]//Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. ACM, 2015: 161-170.