Research on Dynamic Reconfigurable Convolutional Neural Network Accelerator

Bao Deng¹,², *, Hao Lv¹,²
¹School of Computer Science and Engineering, Northwestern Polytechnical University, Xi’an, China
²Xi’an Aeronautics Computing Technique Research Institute, Aviation Industry Corporation of China, Xi’an, China

*Corresponding author e-mail: dengb@avic.com

Abstract. The hardware implementation of convolutional neural network has the problem of resource limitation, which can be solved by the design of convolutional neural network accelerator based on FPGA dynamic reconstruction. The whole parallel strategy and architecture of CNN accelerator are designed, and the functional modules of CNN are designed based on pipeline. The dynamic reconstruction technology is used to redesign the convolutional neural network accelerator, and the dynamic reconstruction region and division are established; BPI flash is selected to store the configuration file, and the configuration file is read internally to dynamically configure the dynamic reconstruction area. Finally, for lenet. 5 handwriting recognition, compared with the corresponding static design, the use of slice LUTS, slice registers and DSP resources of the accelerator based on dynamic reconstruction design is reduced by 46%, 25% and 68% respectively. Compared with the implementation based on software platform. The system execution time is greatly reduced. However, due to the bandwidth limitation of the internal configuration port, the reconfiguration configuration time prolongs the execution time of the whole convolutional network.

Keywords: Dynamic reconfiguration; CNN; Resources; Time; FPGA.

1. Introduction
Convolutional neural network is a kind of multilayer neural network, which is usually composed of convolution layer, pool layer and full connection layer. Convolutional neural network has high parallelism. The operation of convolutional neural network is composed of a large number of multiplication and addition, which is a network with high operation density. FPGA design is flexible, a large number of on-chip resources can meet the requirements of convolutional neural network operation, not only can give full play to the parallelism of convolutional neural network, but also effectively reduce the design cost.

The dynamic reconstruction technology based on FPGA is to divide the configuration area of FPGA into static region and reconstruction area by using the idea of time division multiplexing. Under the control of reconfiguration controller, the reconstruction area is added from memory. After loading
different configuration files, after a certain configuration time, the reconfiguration module is loaded, and then the reconfiguration area can start to perform different logical functions. If the configuration time of the refactoring area is greater than the time interval between the two refactoring modules, the execution time of the system will be prolonged, otherwise, it can be hidden in the execution time. When a reconfiguration area is reconfigured, the normal operation of other areas in FPGA is not affected. By using the characteristics of FPGA reconfigurable dynamically, the time division multiplexing of resources on FPGA can be realized, and some modules are dynamically reconstructed, which improves the flexibility of the system, makes full use of FPGA hardware resources and saves costs. Dynamic reconstruction technology has a high expansibility, and can be applied in many fields such as space, national defense and so on.

2. Basic principles

2.1. Convolutional neural network
Convolutional neural network is usually composed of convolution layer, pooling layer and fully connected layer. Convolution layer and pooling layer are alternately connected to form the first layers of convolution neural network, and one or several fully connected layers constitute the first layer of convolution neural network.

The latter layer is used to identify and classify the features generated by the former network layer. By sharing the weights of local connections and neurons, the parameters of convolutional neural network can be greatly reduced, and the execution efficiency of convolutional neural network is higher than that of full connected neural network. Combined with the function of pooling layer, the image features are invariant to translation, scaling and distortion. The optimized lenet-5 handwriting recognition convolutional neural network used in this paper is a typical multi-layer perceptual network [1], which is composed of three convolution layers (C1, C3, C5), two pooling layers (S2, S4) and a full connection layer (F6), as shown in Figure 1.

![Figure 1. Lenet-5 model structure](image)

2.2. FPGA dynamic reconfiguration
FPGA based dynamic reconfigurable technology is to use the idea of "time division multiplexing" to transform the design from a pure space digital logic system to a digital logic system constructed in time and space. This technology makes. The utilization of FPGA resources has been doubled, and the scale of hardware used to realize system functions has been greatly reduced. As shown in Figure 2.
Dynamic reconfiguration process needs to be controlled by dynamic reconfiguration controller. The dynamic reconfiguration controller receives the trigger signal generated in the static region, and then reconfigures the corresponding reconstruction area. The dynamic reconfiguration controller retrieves some bit stream files from nonvolatile memory (such as flash) and transmits them to the internal configuration access port (ICAP). Dynamic reconfigurator can be placed in external devices (such as processors) or FPGA devices that need to be refactored. Like other logic in static region, dynamic reconfiguration controller is continuously used in the whole dynamic reconstruction process function.

Vivado Design Suite supports the Dynamic Reconfiguration Design in software. In vivado, it provides the special process of Dynamic Reconfiguration. It also provides the partial reconfiguration controller (PRC) IP core [2]. The design of dynamic reconstruction process is carried out.

3. Hardware design

3.1. Static accelerator
The overall hardware architecture of the static convolutional neural network accelerator is shown in Figure 3. In the first layer (C1), the third layer (C3) and the fifth layer (C5), six, six and ten convolution modules are used respectively. Operation. The second pooling layer (S2) and the fourth pooling layer (S4) are pipelined with C1 and C3 respectively. The last layer, F6, directly multiplexes 10 convolution modules of C5 layer.

In the first convolution layer (C1), the input is $32 \times 32$ input image, the output is 6 channel output feature map. So in C1, six convolution modules are designed to calculate six output characteristic maps.
in parallel count. There are five PE units in each convolution module, which decompose 25 multiplication and addition operations into five l units \( \times 5 \).

C1 is followed by the first pooling layer (S2), and the input is 6 pieces of \( 28 \times 28 \), output 6 \( 14 \times 14 \). So in S2, six pooling modules are designed to pool six input feature maps in parallel. C1 and S2 adopt pipeline structure. In S2 phase, the pixels generated in C1 phase are buffered first, and when the buffered pixels can form a \( 2 \times 2 \), the operation of S2 layer is started.

One of the core operations of convolution neural network is image convolution. Because the convolution operation scale of volume layer [3] in lenet-5 network is \( 5 \times 5 \). In this paper, the convolution operation is divided into five equal groups, and the five multiplication and addition operations in each group adopt two-stage pipeline structure. Each group of multiplication and addition operations will output a result to the next group. The next group will add its own multiplication and addition results to the results of the previous group, and then continue to input the next group. Finally, the fifth group accumulates the sum of multiplication and addition of the first four groups for output. When the PE works, the multiplier has input data in each cycle, which not only improves the throughput, but also improves the utilization of hardware.

3.2. Dynamic accelerator

The dynamic part reconstruction scheme designed in this paper divides the FPGA chip into static region and four dynamic reconstruction regions (PRI, PR2, PR3, PR4). Each refactoring area is allocated with three different refactoring modules for time sharing configuration. When the reconfiguration module is allocated to different reconstruction areas, the principle of the same reconfiguration area is followed. The number of on-chip resources occupied by the reconfiguration module should be as close as possible to reduce the waste of on-chip resources. The allocation of reconfiguration module is shown in Table 1. In different stages, the corresponding reconfiguration modules are configured into the corresponding reconstruction area, and the operation of this stage is completed together with other modules in the static region.

| Table 1. Restructuring allocation |
|----------------------------------|
| stage   | PR1       | PR2       | PR3       | PR4       |
| C1-S2   | S2 FIFO   | C1        | Tanh1     | Tanh2     |
| C3-S4   | C3 1      | C3 2      | C3 3      | C3 4      |
| C5-F6   | C5 11     | C5 2      | C5 3      | Tanh5 4   |

For the control of the reconstruction process, the scanning and extraction of configuration files, and the storage of configuration files. When the reconfiguration area needs to be reconfigured, a specific trigger signal will be generated. After capturing the trigger signal, PRC extracts the configuration file stored in BPI flash through ICAP port. ICAP, as the internal configuration port of FPGA, has the same function as El in selectmap [4]. It will dynamically configure FPGA under the control of PRC.

4. Test and analysis

In this paper, the Dynamic Reconfiguration Design of lenet-5 handwriting recognition network is implemented on xilinx.xc7z020 development board. Verilog hardware description language is used to write the code, and vivado ide 2018 is used for simulation. The software environment is windows 10.

4.1. Resource assessment

The static design (without dynamic part reconstruction) of lenet-5 convolutional neural network is compared with that of dynamic part reconstruction. The dynamic part reconstruction method is used to design on chip resources compared with static design. The usage of DSP, especially DSP, has been greatly reduced. The specific evaluation results are shown in Table 2.
Table 2. Resource occupancy

| Design method     | Slice LUT | Slice Register | DSP |
|-------------------|-----------|----------------|-----|
| Static design     | 41825     | 53925          | 570 |
| Dynamic design    | 23626     | 38932          | 157 |
| %                 | 55.5      | 71.2           | 27.9|

After the dynamic reconfiguration design, the slice LUTs, slice register and DSP used in the CNN accelerator are reduced to 55.5%, 55.5%, 71.2% and 27.9% of the static design respectively. It can be seen that the dynamic reconfiguration design can effectively reduce the hardware resources occupied by the convolutional neural network accelerator, improve the resource utilization, and save the hardware cost.

4.2. Time assessment

The configuration speed of FPGA dynamic reconfiguration is limited by the size of configuration file and the bandwidth of ICAP port [5].

In this paper, there are four reconstruction areas, each of which has three time-sharing reconfiguration modules. Therefore, a complete handwriting recognition network requires a static configuration file and 12 parts sub configuration file. The total profile size is about 27MB.

XC7Z020 series development board, the maximum bandwidth of ICAP configuration port at 100 MHz maximum clock frequency is 3.2GB/s. The total time is 20.3ms, which is calculated by the recognition time of convolution neural network. Compared with the software implementation based on arm cortex A9 processor, as shown in Table 3, there is a great improvement in time and power consumption.

Table 3. Identify one hundred consumption comparisons

| platform          | the way          | Time/ms | Power/W |
|-------------------|------------------|---------|---------|
| ARM Cortex A9     | software         | 5010    | 45      |
| Xilinx XC7Z020     | Dynamic reconstruction | 1102   | 7.8     |

Due to the bandwidth limitation of the internal configuration port of ICAP [6], the time required to complete a handwriting recognition after dynamic reconstruction is longer than that for static implementation on FPGA. This is because in the refactoring module needs to work, the refactoring area has not been reconfigured. The time of waiting for the refactoring area to be configured prolongs the execution time of the whole system.

5. Conclusion

This paper introduces the idea and design method of dynamic partial reconstruction, and designs the convolutional neural network of Lenet.5 handwriting recognition. The resources, time and power consumption of the design on Xilinx xc7z020 are given. The experimental results show that the dynamic reconfiguration design can effectively reduce the on-chip resources needed in the hardware implementation phase of convolutional neural network, improve the utilization of hardware resources, and make the hardware implementation of convolutional neural network more flexible.

In this paper, the dynamic reconstruction technology is applied to the hardware implementation of convolutional neural network, which can effectively solve the problem that convolutional neural network can’t be completely placed in the hardware, because it needs too many hardware resources [7]. The bandwidth of ICAP configuration port is the main limiting factor of dynamic reconfiguration configuration speed. We hope that future research work can improve this.

References
[1] Lu Lu. Study of convolutional neural network and its applications on license plate
recognition. Hefei: Hefei University of Technology, 2006.

[2] Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009: 248-255.

[3] Xilinx. Vivado design suite tutorial: partial reconfiguration (ver2018.2). [2018-12-21]. https://china.xilinx.com/support/documentation/sw-manuals/xilinx2018-2/19947-vivado-partial-reconfiguration-tutorial.pdf.

[4] Xilinx. Partial reconfiguration controller product guide (ver4.5). [2018-12-21]. https://china.xilinx.com/support/documentation/ip-documentation/prc/vl-3/p9193-partial-reconfiguration-controller.pdf.

[5] Xilinx. Partial reconfiguration user guide (ver4.5). [2018-12-21]. https://china.xilinx.com/support/documentation/sw-manuals/xilinx 147/u9702.pdf.

[6] Xilinx. BP1 fast configuration and MPACT flash programming with 7 Series FPGAs application note (ver1.2). [2018-12-21]. https://china.xilinx.com/support/documentation/application-notes/xapp587-bpi-fast-configuration.pdf.

[7] He Kaixuan, Yuan Xun, Chen Song. Design of convolutional neural network hardware architecture based on FPGA dynamic reconfiguration. Information Technology and Network Security. China. 2019. Vol38.