Research on the Overview of Image Processing Architecture of Computer Based Deep Neural Network Accelerator

Deqin Shu¹, Hao Fan¹*, Liang Zhang¹
¹College of Information Science & Engineering, Shandong Agricultural University, China

*Corresponding author e-mail: fanhao@sdau.edu.cn

Abstract. DNN algorithm still has many shortcomings in the process of operation, which need to be further solved. Specifically, there is more data reuse, and the repeated access of global cache takes up more resources and computation, thus reducing the efficiency of operation. Based on this, this paper first analyses the research status and value of the DNN accelerator, then studies the image processing architecture of the DNN accelerator, and finally gives the computer DNN model and acceleration algorithm analysis.

Keywords: Image Processing Architecture, DNN, Accelerator

1. Introduction
With the iterative progress and maturity of computer tech, it has been widely and deeply studied and popularized in many fields, especially the utilization of computer tech represented by DNN accelerator in image processing architecture, which greatly accelerates the amelioration of image processing level [1]. DNN can obtain high-dimensional features from a large number of letters, including several aspects as shown in Figure 1 below, and can obtain high recognition accuracy. With the more and more intelligent and info utilization scenarios, not only the scene complexity of image processing is also increasing, but also the amount of calculation and data is growing significantly, which need to further ameliorate the accuracy of recognition and processing.

![Figure 1. High dimensional features that can be obtained by deep neural network](image-url)

DNN algorithm still has many shortcomings in the process of operation, which need to be further solved. Specifically, there are more data reuse, and the repeated access of global cache takes up more resources and computation, thus reducing the efficiency of operation. In view of these shortcomings of
DNN algorithm, the function of accelerator is fully exploited to ameliorate the data reuse rate and reduce the problem of repeated transmission. At present, most of the DNN accelerators also have a lot of room for optimization and amelioration, which are manifested in many aspects, such as large repeated access of feature map data, low utilization rate of processing unit, complex control and so on. In this context, it is necessary to develop an accelerator image processing architecture that can satisfy the DNN algorithm and realize the effective reconstruction of computing array and mapping method.

In addition, the structural design of DNN accelerator focuses on the neuron circuit structure, link organization and data flow control [2]. DNN algorithm has made remarkable achievements in the field of unstructured data processing represented by image recognition, classification and processing, and structured data processing represented by advertisement classification and search optimization. With the complexity and diversification of utilization scenarios, DNN algorithm's requirements for hardware computing power are also increasing. Utilization accelerator can ameliorate its computing efficiency, especially the data reuse rate of DNN.

In short, aiming at the typical characteristics of DNN algorithm, such as post-processing calculation, large amount of data, multiple loops, and sparsity of calculation and so on, the accelerator is used to optimize the design and dynamic adaptation of the algorithm, so as to ameliorate the computational power and efficiency of the computing unit [3]. With the continuous optimization of data set and DNN model, DNN accelerator can further ameliorate the applicability and matching of its utilization field. Therefore, it is of great value to study the image processing architecture of computer-based DNN accelerator.

2. Research status and value of DNN accelerator

2.1. Research status and trend of computer DNN accelerator
With the diversification of DNN utilizations, the task scenarios are becoming more and more complex [4]. In order to reduce the excessive dependence on hardware computing power, we need to integrate accelerators in the architecture of DNN to realize the reasonable balance of efficiency and computing power, so as to greatly expand the utilization scope and scene of DNN algorithm. At present, the research of DNN accelerator mostly focuses on the implementation of algorithm and calculation, but with the gradual dominance of large-scale features of advanced neural network, the design of DNN accelerator also needs to pay more attention to its performance and energy [5]. High throughput accelerators can use advanced machine learning algorithms to gain more extensive and systematic utilizations.

2.2. The value of DNN accelerator
With the evolution of DNN architecture, how to design a hybrid core accelerator to balance the two key indicators of flexibility and efficiency has become the focus and difficulty of research. The architecture of DNN is gradually transiting to accelerator.

![Figure 2. Research fields of DNN accelerator](image-url)
In the development of high-performance and embedded utilizations, including several aspects as shown in Figure 2 above, accelerator design based on neural network can achieve high performance and efficiency. The accelerator focuses on image processing, and studies the architecture and control of the accelerator to minimize memory transfer and execute as effectively as possible [6]. Inefficient memory transfer may result in inefficient throughput, accelerator has energy or cost advantages, and memory transfer in the target algorithm is of specific nature. Expand the scale of the neural network to achieve better accuracy and more functions.

2.3. Design points of DNN accelerator
The training of DNN relies on back propagation algorithm, which inputs eigenvectors from the input layer and calculates the output layer by layer. The output layer finds that the output is different from the correct class number. At this time, it allows the last layer of neurons to adjust the parameters. The last layer of neurons not only adjusts the parameters themselves, but also orders the second layer of neurons connecting it to adjust, and the layers are adjusted backward. The adjusted network will continue to test on the sample [7]. If the output is still wrong, continue to adjust it in turn until the network output is satisfied. Design of large CNN and DNNS accelerators for synthetic layout and wiring, advanced machine learning algorithm, the accelerator achieves high throughput in small area, power and energy consumption. In addition, the accelerator is designed to focus on memory behavior, and measurements are not affected by computing tasks, these factors in performance and energy storage transfer.

3. Image processing architecture of DNN accelerator

3.1. The necessity and types of DNN acceleration
With the diversification of DNN utilization scenarios and the amelioration of model performance, the power consumption of DNN model is higher and higher, which gradually limits its further in-depth utilization in specific environment and real-time scene. Secondly, the storage cost of DNN model is also increasing, and the memory and calculation times for image processing are also increasing significantly [8]. This is the DNN in the computing; storage capacity of weak terminal equipment is difficult to effectively run. In addition, the parameters and structure of DNN model are not directly related to the discriminant performance. By accelerating the original DNN model, it can be better applied to remote utilization scenarios and requirements.

Common methods of DNN acceleration include several types as shown in Table 1 below. These methods can effectively complete CNN acceleration tasks.

| Methods                   | Functions                                      | Features                          |
|---------------------------|------------------------------------------------|-----------------------------------|
| Parameter pruning         | Remove redundant parameters                    | Combined with low rank             |
| Parameter sharing         | Explore the redundancy of model parameters     | Flexible and effective             |
| Low rank decomposition    | Original convolution kernel in decomposition   | Flexible and effective             |
| Design of compact         | Reduce storage and computing                   | Dependence training model          |
| convolution kernel        | complexity                                      |                                   |
| Knowledge distillation    | Knowledge transfer to compactness model         |                                   |

3.2. Mechanism optimization of DNN accelerator system
In order to achieve high efficiency ratio, the image processing architecture of the accelerator based on DNN needs to be designed in the selection of the number of operational components, organization strategy and the structure parameters of ram on chip [9]. In the whole structure parameter space of
accelerator image processing system, we try to try various possible design parameters, carry out processor performance modeling, select various design parameters, so as to realize the balance between operation and memory access, and further ameliorate the performance of the DNN algorithm. In addition, in order to further reduce the demand of hardware calculation and power consumption of DNN, further optimization of data handling process is needed. By dividing the DNN into blocks, the data handling times needed for neural network operation are optimized.

4. Computer DNN model and accelerated algorithm

4.1. Types and characteristics of DNN models

There are many kinds of DNN models, each model has different network structure, and the difference of network structure is mainly reflected in the number, type, internal connection and shape of layers. Common DNN models are shown in Table 2 below. The depth and width of these models are also increasing with the amelioration of CNN performance. However, the complexity of convolution layer floating-point computing and the continuous amelioration of computing power also make CNN acceleration more and more important.

| DNN models | Convolution layer | Fully connected layer | Feature |
|------------|-------------------|-----------------------|---------|
| LaNet      | 2                 | 2                     | Apply to check deposit task |
| AlexNet    | 5                 | 5                     | Ameliorate classification accuracy |
| VGGNet     | 13                | 3                     | Strong expansibility |
| GooLeNet   | 3                 | 1                     | Reduced model storage time |
| ResNet     | 1                 | 1                     | With 16 residual blocks |

4.2. Accelerated algorithm of parameter pruning for computer DNN

The acceleration algorithm of computer DNN based on network parameter pruning can remove the redundant weight value of DNN model and reduce the model parameters, so as to ameliorate the computational efficiency and reduce the occupation of storage space [10]. Secondly, through the over fitting of pruning network prevention model, pruning is divided into structured and unstructured forms. The former directly delete the whole filter structured info, while the latter delete the parameters of the elements in the filter. In addition, in the training phase of pruning acceleration algorithm, the learning of important connections is realized, and in the deletion phase, the norm of weight matrix is calculated to establish a sparse network. In the phase of retraining weights, the recognition accuracy of the network is recovered.

By selecting the optimization function and deleting the redundant channels, the DNN network stack layer can be deleted, which can speed up the network computing process without increasing the data type storage.

4.3. Other accelerated algorithm of computer DNN

Firstly, the computer DNN acceleration algorithm based on parameter sharing can greatly reduce the memory and computation times, and optimize the regularization order of convolution kernel. Secondly, the DNN acceleration algorithm based on low rank decomposition uses the convolution kernel in the matrix decomposition depth model. The convolution calculation of the low rank factor is shown in Figure 3 below. The low discriminative and significant info is removed from the input calculation space and channel redundancy. In addition, the DNN acceleration algorithm based on the compact convolution kernel accelerates the calculation of the network and achieves higher performance of image target recognition. DNN acceleration based on knowledge distillation can greatly reduce the computational overhead.
5. Conclusion
In summary, with the continuous optimization of data set and DNN model, DNN accelerator can further ameliorate the applicability and matching of its utilization field. Based on the research status and value of DNN accelerator, this paper analyzes the design points of DNN accelerator. Through the analysis of the image processing architecture of the DNN accelerator, the optimization of the DNN accelerator architecture is studied. Finally, the computer DNN model and acceleration algorithm are studied.

Acknowledgments
This work is supported by Natural Science Foundation of Shandong Province (ZR2016FB19, ZR2019MF026 and ZR2017MD018), National Natural Science Foundation of China (41471299), Technology Innovation Fund Project of Shandong Agricultural University (23659).

References
[1] Luo Shuyan, Zhu Yian, Wang Weichao. Research on energy saving capability measurement model of embedded system [J]. Journal of Northwest University of tech, 2016, 34 (03): 536-543.
[2] Pian Zhaoyu, Shi Tianyu, yuan depeng. Utilization of hierarchical visual feature perception in target recognition [J]. Journal of computer aided design and graphics, 2017, 29 (06): 1093-1102.
[3] Qiao Ruixiu, Chen Gang, Gong Guoliang. A high performance reconfigurable deep convolution neural network accelerator [J]. Journal of Xi'an University of Electronic Science and tech, 2019 (3): 38-41.
[4] Shan Shiguang, Liu Xin. Deep learning: the revival and change of multi-layer neural network [J]. Science and tech guide, 2016, 34 (14): 60-70.
[5] Tian Zheng, duhuimin, Huang Xiaokang. Ameliorated piecewise linear approximation method of transcendental function [J]. Computer utilization, 2016, 36 (07): 1807-1810.
[6] Wang Kangkang, Wang Keyan, Li Yunsong. Refocusing algorithm of small text image based on convolutional neural network [J]. Journal of Xi'an University of Electronic Science and tech, 2018, 45 (4): 80-85.
[7] Wu Kai, Chen Yunfang. Embedded experimental platform based on open source virtual desktop [J]. Computer education, 2017 (09): 154-157.
[8] Wu Liangjing, Cao Yunfeng; Ding Meng. Design of visual algorithm acceleration system based on socfpga [J]. Utilization of single chip microcomputer and embedded system, 2016, 16 (11): 58-62.
[9] Zhai Junhai, Zhang Sufang, Hao Pu? Convolutional neural network and its research progress [J]. Journal of Hebei University, 2017, 37 (06): 640-651.
[10] Zhao Bo Ran, Zhang Li, Shi Guang Ming, et al. Design of programmable neural network processor with transmission trigger architecture [J]. Journal of Xi'an University of Electronic Science and tech, 2018, 45 (4): 92-98.