Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 2, 2023

Hardware Accelerated Optimization of Deep Learning Model on Artificial Intelligence Chip

Zhimei Chen

University of California, Santa Barbara, USA

Abstract: With the rapid development of deep learning technology, the demand for computing resources is increasing, and the accelerated optimization of hardware on artificial intelligence (AI) chip has become one of the key ways to solve this challenge. This paper aims to explore the hardware acceleration optimization strategy of deep learning model on AI chip to improve the training and inference performance of the model. In this paper, the method and practice of optimizing deep learning model on AI chip are deeply analyzed by comprehensively considering the hardware characteristics such as parallel processing ability, energy-efficient computing, neural network accelerator, flexibility and programmability, high integration and heterogeneous computing structure. By designing and implementing an efficient convolution accelerator, the computational efficiency of the model is improved. The introduction of energy-efficient computing effectively reduces energy consumption, which provides feasibility for the practical application of mobile devices and embedded systems. At the same time, the optimization design of neural network accelerator becomes the core of hardware acceleration, and deep learning calculation such as convolution and matrix operation are accelerated through special hardware structure, which provides strong support for the real-time performance of the model. By analyzing the actual application cases of hardware accelerated optimization in different application scenarios, this paper highlights the key role of hardware accelerated optimization in improving the performance of deep learning model. Hardware accelerated optimization not only improves the computing efficiency, but also provides efficient and intelligent computing support for AI applications in different fields.

Keywords: Artificial Intelligence Chip; Hardware Accelerated; Deep Learning.

1. Introduction

In recent years, with the rapid development of artificial intelligence (AI) technology, deep learning model has become the key driving force for breakthroughs in many application fields. However, with the increasing complexity and scale of the deep learning model, the demand for computing resources has also increased exponentially. In order to meet this demand, researchers have turned their attention to hardware acceleration technology, among which AI chip has attracted much attention as an important hardware implementation method.

The rise of AI chip indicates that hardware design is moving towards a more specialized and efficient direction. Compared with the traditional general-purpose processor, AI chip can provide higher energy efficiency ratio and computing performance, so that the deep learning model can complete the training and inference tasks in a shorter time [1]. In the research of hardware accelerated optimization of deep learning model, a problem that cannot be ignored is how to better balance the requirements of computing and storage. The training and inference process of deep learning model need a lot of parameters and intermediate data, so the efficient use of storage system becomes a key aspect to improve performance. The focus of this paper is on the hardware acceleration optimization of deep learning model on AI chip, aiming at deeply discussing how to achieve efficient processing of deep learning tasks through clever design and optimization of hardware structure.

Through in-depth study of hardware acceleration technology, this study is expected to provide strong guidance for the design and development of AI chips in the future and promote the wide application of deep learning technology in various fields. The research results of this paper will provide substantial support for improving the performance of deep learning model, reducing energy consumption and promoting the innovative development of AI technology.

2. Deep Learning Model

With the rapid development of deep learning, convolutional neural network (CNN), as a deep learning model specially used for processing images and sequence data, has gradually become one of the core technologies in computer vision, natural language processing and other fields [2-3]. The success of CNN lies not only in its excellent performance, but also in its ability to automatically learn and express data features hierarchically. The model structure of the classic CNN model AlexNet is shown in Figure 1.

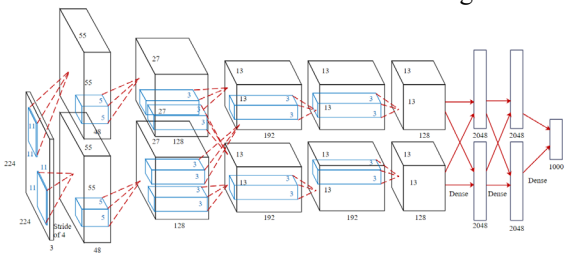

Figure 1. Model structure of AlexNet

First of all, the core idea of CNN is to capture local features in input data through convolution operation. This way of local perception enables CNN to efficiently model the spatial hierarchy in the image, so as to better understand and extract the features of the image [4]. By sharing weights, the convolution layer effectively reduces the number of parameters, improves the calculation efficiency of the model, and realizes the modeling of translation invariance while maintaining sensitivity to local features, making the model more generalized.

Secondly, the introduction of pooling layer further enhances CNN's abstract ability. Through pooling operation, CNN can reduce the dimension of data while retaining important information, accelerate the training process of the model, and reduce the risk of over-fitting. The role of pooling layer is to extract the main features, so as to achieve a more abstract expression of images or sequence data at a high level.

The multi-level structure of CNN also makes it suitable for feature extraction at different levels. From the bottom edge and texture to the high-level semantic information, CNN has formed an automatic learning feature extractor through feature extraction and abstraction layer by layer, which makes the model adapt to the feature requirements of different tasks [5].

In addition, by introducing activation function and regularization technology, CNN can better deal with nonlinear relations and prevent over-fitting, and improve the generalization ability of the model. In recent years, the deep learning community has continuously improved and optimized CNN, and introduced innovative structures such as residual network and attention mechanism, which made CNN achieve remarkable results in more complex tasks and larger data sets.

As a powerful deep learning model, CNN has not only performed well in the field of image processing, but also achieved remarkable success in many other fields [6-7]. Its hierarchical feature learning and sensitivity to local information make CNN have a wide application prospect in dealing with complex data and solving practical problems.

3. Analysis of Hardware Characteristics of AI Chip

With the continuous progress of AI technology, AI chip is the core hardware component to support and promote AI application, and the analysis of its hardware characteristics is very important for understanding its performance and application fields. AI tasks usually involve large-scale matrix operations and neural network training, which puts great demands on computing power. One of the hardware characteristics of AI chip is its powerful parallel processing ability. Through parallel computing unit and specific hardware structure, AI chip can process multiple data at the same time, improve computing efficiency, and thus accelerate the training and inference process of deep learning model [8].

Energy efficiency is one of the most important indicators in AI chip design. Compared with traditional general-purpose processors, AI chips usually adopt special architecture and design to improve energy efficiency. Hardware-specific optimization, such as hardware acceleration and low-power design of deep learning algorithm, enables AI chips to use energy more efficiently and reduce computing costs when performing large-scale AI tasks.

The design of AI chip usually integrates a special neural network accelerator. These accelerators are hardware units for deep learning tasks, which accelerate the forward propagation and backward propagation of neural networks by efficiently performing convolution and matrix operations. This hardware feature makes AI chip have excellent performance when dealing with large-scale neural networks. Although AI chips usually have special hardware accelerators, their design also pays attention to flexibility and programmability. This means that AI chips can adapt to different AI tasks and be used flexibly in different application scenarios [9]. The programmable design enables the chip to adapt to the new algorithm and model structure through firmware or software update.

In order to improve the performance of the whole system, AI chips are usually highly integrated. In addition to the neural network accelerator, it is also possible to integrate multiple functional units such as storage controller and memory to reduce communication delay and improve data throughput, thus better supporting large-scale deep learning tasks. In order to better adapt to different types of AI tasks, AI chips usually adopt heterogeneous computing structures. This design allows different types of computing units to work together to optimally complete various complex AI tasks, including image recognition, natural language processing and so on.

The hardware characteristics of AI chip show remarkable advantages in its parallel processing ability, energy-efficient computing, neural network accelerator, flexibility and programmability, high integration and heterogeneous computing structure. These characteristics provide powerful computing support for AI applications and promote the rapid development of AI technology in various fields. With the continuous progress of technology, the hardware characteristics of AI chips will continue to evolve, providing more efficient and innovative solutions for a wider range of application scenarios.

4. Hardware Accelerated Optimization of CNN on AI Chip

With the rapid development of deep learning technology, CNN, as the core model in tasks such as image processing and pattern recognition, is widely used in AI applications, which puts higher demands on computing resources. In order to give full play to the potential of CNN on AI chip, hardware accelerated optimization has become an inevitable choice to improve performance and efficiency.

The training and inference process of CNN involves a lot of convolution and matrix operations, which puts great demands on computing resources. Through hardware acceleration optimization on AI chip, we can use hardware parallelism and special convolution accelerator to significantly improve the computational efficiency. This helps to shorten the training time of the model and accelerate the real-time inference task, making the AI application more responsive.

By using CNN on AI chip to accelerate hardware optimization, we can improve computing efficiency, reduce power consumption, achieve higher real-time performance, and support complex models and large-scale data processing. This not only has an important impact on the performance improvement of current AI applications, but also opens up new possibilities for broader AI application scenarios in the future [10].

When designing CNN for hardware acceleration optimization on AI chip, we need to consider many factors such as hardware structure, computing efficiency and energy consumption. A special convolution accelerator is designed, which can perform convolution operations in parallel to improve the computational efficiency. The calculation formula of convolution operation is:

$$C_{ij} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left(A_{i+m, j+n} \times M_{m,n} \right) \quad (1)$$

Where C_{ij} is the element of output feature map, $A_{i+m,j+n}$ is the element of input feature map, and $M_{m,n}$ is the weight of convolution kernel.

Using distributed cache structure, the input data and model parameters are quantized, and floating point numbers are converted into fixed points, which reduces data storage requirements and speeds up calculation. The calculation formula of quantification is:

$$Quantized\_Data=round\left(\frac{Float\_Data}{Scale\_Factor}\right) \quad (2)$$

In the direct convolution optimization, there are several levels of data sharing. In the original algorithm, one thread completes the calculation of one neuron, so each thread needs to input 3×3 inputs, and 16 neurons need a total of 16× 3 inputs. Assuming that one thread completes the calculation of 16 neurons, it only needs to input 6×6 inputs, which is 1/4 of the original input. If different worker threads in the same wrap can share data, as long as there is enough filter, the input can be greatly reduced.

5. Experimental Analysis

In order to verify the applicability and effectiveness of the proposed software-hardware collaborative framework in CNN optimization, this section embeds the framework into the mainstream CNN model GooleNet. In this paper, cuDN, a deep neural network library developed by NVIDIA, is used as the experimental comparison baseline to realize the optimization and acceleration of GoogleNet. On the other hand, the framework proposed in this paper is embedded into the network architecture of GoogleNet, which is converted into matrix multiplication as the input of the kernel function batch matrix, and then the matrix is processed by the proposed batch GEMM kernel function in parallel to realize the parallel optimization of convolution, thus accelerating the training process of GoogleNet.

Figure 2 shows the time performance comparison of inception3a layer in GoogleNet. It can be seen that the overall running time of the inception3a layer after batch processing is lower than that of cuDNN, and the average performance acceleration is about 1.26x.

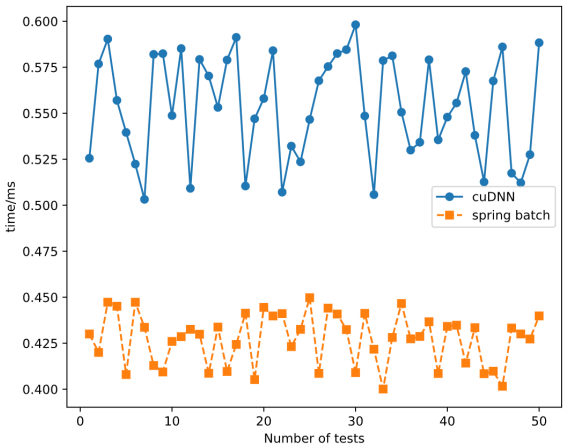

Figure 2. Comparison of time performance of inception3a layer

Therefore, experiments show that embedding the end-to-end collaborative processing framework into CNN model can improve the performance, and the framework proposed in this paper has its applicability for networks with large convolution operations, and can operate batch matrices. The experimental results show that the proposed framework accelerates the performance of CNN model compared with cuDNN, which is of great significance to speed up the training process of CNN.

6. Conclusion

By comprehensively analyzing the key elements of hardware acceleration optimization, including parallel processing ability, energy-efficient computing, neural network accelerator, flexibility and programmability, high integration and heterogeneous computing structure, this paper deeply discusses the method and practice of optimizing deep learning model on AI chip. The design and optimization of neural network accelerator has become the core of hardware acceleration. Through the special hardware structure, the efficient acceleration of deep learning calculation such as convolution and matrix operation are realized, which provides strong support for the real-time performance of the model. Through the hardware accelerated optimization of deep learning model on AI chip, we can not only better solve the bottleneck problem of traditional computing platform, but also provide more powerful and efficient computing support for AI applications in emerging fields. Hardware accelerated optimization on AI chip is one of the key driving forces for the continuous evolution of deep learning technology. Through in-depth research and innovative design of hardware structure, we are expected to further promote the development of AI technology and realize a more intelligent, efficient and widely used AI system.

References

[1] Gupta, S. , Nguyen, D. , Rana, S. , Venkatesh, S. , & Kuttichira, D. P. (2022). Verification of integrity of deployed deep learning models using bayesian optimization. Knowledge-based systems (6), 241.

[2] Khalifa, N. E. M., Taha, M. H. N. , Ali, D. E. , Slowik, A. , & Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor rna-seq data: a novel optimized deep learning approach. IEEE Access(8), 8.

[3] Singh, K. , & Kapania, R. K. (2021). Accelerated optimization of curvilinearly stiffened panels using deep learning. Thin-Walled Structures, 161(3), 107418.

[4] Geng, X. , Zhao, L. , Shi, L. , Yang, J. , & Sun, W. (2021). Small-sized ship detection nearshore based on lightweight active learning model with a small number of labeled data for sar imagery. Remote Sensing, 13(17), 3400.

[5] As, I. , Pal, S. , & Basu, P. (2018). Artificial intelligence in architecture: generating conceptual design via deep learning. International Journal of Architectural Computing, 16(4), 306-327.

[6] Aljohani, T. M. , Ebrahim, A. , & Mohammed, O. (2021). Real-time metadata-driven routing optimization for electric vehicle energy consumption minimization using deep reinforcement learning and markov chain model. Electric Power Systems Research(6), 192.

[7] Asfahan, H. M. , Sajjad, U. , Sultan, M. , Hussain, I. , & Khan, M. U. (2021). Artificial intelligence for the prediction of the thermal performance of evaporative cooling systems. Energies, 14(13), 3946.

[8] Ropiak, K. , & Artiemjew, P. (2020). On a hybridization of deep learning and rough set based granular computing. Algorithms, 13(3), 63.

[9] Dong, X. , Dong, C. , Chen, Z. , Cheng, Y. , & Chen, B. (2020). Botdetector: an extreme learning machine-based internet of things botnet detection model. Transactions on Emerging Telecommunications Technologies (4), e3999.

[10] Xia, J., Deng, D. , & Fan, D. (2020). A note on implementation methodologies of deep learning-based signal detection for conventional mimo transmitters. IEEE Transactions on Broadcasting, (99), 1-2.