Dean&Francis

ISSN 2959-6157

Design and Optimization of VLSI Architectures for Artificial Neural Networks: Innovations, Challenges, and Future Directions

Anyongyong Zhao

Case Western Reserve University, Cleveland, OH, United States of America

axz380@case.edu

Abstract:

With the rapid development of artificial technology, artificial neural networks (ANN) have come into people's vision. Artificial neural network is a computing system that mimics the workings of the human brain. It is widely used in information recognition, natural language processing and predictive analysis. It uses electronic components to form nodes. These nodes are similar to human neurons: each node is connected with multiple nodes in adjacent levels with different weights. Very large scale integrated circuit (VLSI) technology plays a crucial role in optimizing the hardware implementation of ANN. This paper introduces the design principle of VLSI architecture for ANN. Using a simple ANN model- Perceptron as an example, the optimization design of tree adder and accumulator is discussed. In addition, several advanced design techniques and hardware-specific optimization strategies of VLSI technology in practical ANN applications will be discussed. Lastly, this paper summarizes the challenges that VLSI may face in the development of neural networks.

Keywords: VLSI; artificial neural network; perceptron.

1. Introduction

Science and technology serve human beings, and only by understanding human thinking mode and thought structure can we better create value. In the artificial intelligence community, scientists are interested in combining artificial intelligence and the human brain: having technology mimic the way the human brain processes large amounts of data [1]. This technology becomes the key technology to solve complex problems. As technology improves, however, these networks require even more powerful computational support: it requires tens of thousands of tiny electronic components to form a structure similar to the interconnectivity of neurons in the human brain. This structure provides powerful computing support, but it also requires strong technical requirements [2]. This shows the importance of VLSI. VLSI refers to the integration of a large number of transistors on a single silicon chip to achieve complex electronic systems, which perfectly meets the needs of ANN for a large number of micro components. Using

VLSI technology, we can design special hardware for artificial neural networks, such as a neural processing unit (NPU), which can efficiently perform the parallel computing tasks of a neural network [3]. In addition, VLSI allows for more compact designs and lower power consumption, which is particularly important for neural network systems that require long running or embedded applications. By optimizing circuit design and architecture, VLSI can not only increase the processing speed of ANN, but also reduce the physical space footprint while maintaining high energy efficiency [4]. VLSI has also become the core of modern electronic engineering with its high integration, energy saving, and low cost, playing a key role in the development of computer hardware and electronic devices. By customizing VLSI chips to optimize neural network structures, researchers and engineers can design systems that are better suited to different specific tasks.

2. Technical Overview

2.1 Architecture of Artificial Neural Networks (ANN)

The VLSI architecture is designed with various elements: logic gates, circuits, memory units, and interconnection networks [5]. These elements work together to achieve the desired electronic function. Multiple processing units are designed on the chip to realize highly parallel data processing. ANN usually involves a lot of matrix operations. These calculations are often time-consuming. Each processing unit can perform operations independently, greatly increasing the speed of computation. For complex neural networks, they typically involve millions to billions of parameters and large amounts of and input data. Independent processing units can process different parts of the neural network at the same time, assigned to different inputs, thus maximizing work efficiency. Moreover, by distributing computing tasks to multiple independent processing units, VLSI architectures can reduce overall energy consumption. Each processing unit is responsible for only small computational tasks and can be put into low-power mode when not in use, saving energy when performing complex neural network tasks.

Before the emergence of VLSI, the traditional architecture was the Central processing unit (CPU) or graphics processing unit(GPU) [6]. The VLSI architecture provides ANN with specially optimized hardware resources, which has the following advantages over the traditional architecture. Traditional CPU or GPU are designed for general-purpose computing tasks, and the instructions the processor executes cover a wide range of tasks. For complex and intensive matrix computation, traditional architectures are not as efficient as VLSI architectures in resource utilization. Moreover, traditional architectures are highly dependent on external memory. CPU and GPU often rely on external memory for data access, which increases the latency of data transfer. Especially when dealing with large-scale neural networks, frequent memory access can cause bottlenecks. The on-chip cache and local memory structure in VLSI design greatly reduce the data transmission delay and access time. The efficiency of data transfer between local processing units reduces the dependence on external memory, thus speeding up the computation.

2.2 Design of the Tree Adder

The tree adder is a hardware structure that computes addition operations in parallel. It adds multiple inputs hierarchically. The basic structure of a tree adder is to group multiple input values in the form of a “tree”: first add the input data in pairs to generate a partial sum, and then continue to add these partial sums in pairs until a final sum is obtained [7]. It can be seen that the tree adder has high technical requirements for parallel execution at the same time. VLSI has strong parallel processing capability and is perfectly suited to the requirements of tree adder. By grouping calculations and recursively merging partial sums, it reduces the number of calculations and thus the overall addition delay. This can significantly reduce the delay in feedforward propagation, making the forward computation of the whole neural network more efficient. In backpropagation, errors propagate from the output layer to the hidden layer. In each layer, the errors of the calculated nodes need to be weighted by summing the errors of multiple connections [8]. By means of parallel addition, the tree adder can quickly complete the error accumulation and gradient calculation, and accelerate the whole process of backpropagation. In contrast to traditional adders, which are usually executed sequentially, each addition operation must wait for the previous operation to complete, resulting in O(n) time complexity. The tree adder reduces the time complexity to O (log n) by hierarchical parallel addition, especially in large-scale addition operations, showing significant speed advantages.

2.3 Implementation of the Accumulator

In ANN, the accumulator is one of the essential components of the calculation process. It is responsible for accumulating multiple input values and maintaining the accuracy of intermediate results in multiple operational steps [9]. In the forward propagation stage, each neuron needs to multiply the input value with the corresponding weight and add it up to get the weighted sum, and then generate the output through the activation function.

ISSN 2959-6157

This accumulation process is repeated many times in a multi-layer neural network. Therefore, the accumulator needs to efficiently handle large-scale addition operations to ensure the forward computation speed of the network. With VLSI technology, the accumulator can speed up the computation through parallel processing [10]. VLSI architectures can perform multiple accumulation operations at the same time, which significantly reduces the accumulation delay time. This parallel processing approach is especially suitable for scenarios that require a lot of parallel computation in deep neural networks. The frequent operation of the accumulator generates a large amount of power consumption. VLSI technology can reduce the power consumption of the accumulator by optimizing the logic design of the circuit and using low-power components. By reducing switching activities and optimizing power management, VLSI accumulators can perform operations efficiently while maintaining low power consumption. The accumulator plays an important role in the computation of artificial neural networks, especially in the process of forward propagation, back propagation and weight updating. VLSI technology significantly improves the efficiency of accumulators by means of parallelization design, pipeline processing, low-power logic design and dynamic frequency regulation. In the future design of neural network hardware accelerators, the optimization of the accumulator will still be one of the key factors to improve the performance of the system. Through continuous innovation and optimization, the accumulator will continue to push neural network computing to become more efficient and energy efficient.

3. Perceptron Algorithm

Perceptron is the basic unit in artificial neural networks for classifying linearly separable sample data. The perceptron is a linear classification algorithm that determines the output based on the weight and threshold comparison results of the input. The input to the perceptron is a feature vector: X= [x1, x2…xn], and each input feature xi corresponds to a weight wi. The importance of the input in the decision determines the size of the weight. The weighted sum of the input is calculated by the following formula:

$$Z = \sum_{i=1}^{n} ?w_i \cdot x_i + b \tag{1}$$

The weight update rules of the perceptron are straightforward: each time a prediction is wrong, the perceptron adjusts its weight according to the size of the error. This online learning allows the perceptron to gradually adjust parameters to minimize classification errors until it finds a correct linear separation hyperplane. This rule later inspired gradient descent algorithms. A significant limitation of perceptrons is that they can only solve linearly separable problems, not linearly indivisible data sets.

As the basic unit, the perceptron neural network can be embedded into a deep learning accelerator, and is particularly suitable for processing large-scale perceptron networks, such as multi-layer perceptrons or classification tasks in deep neural networks. In embedded systems, due to limited power consumption and hardware resources, VLSI implementation of perceptron algorithm can provide efficient and low-power classification function, which is widely used in mobile devices, iot sensors and edge computing devices. Through hardware optimization of VLSI, perceptrons can be used in applications that require high real-time and low latency, such as automated driving, industrial automation, and real-time data classification in medical devices. By introducing parallel computing, pipelined design, low-power design and on-chip storage optimization into VLSI architecture, the computational efficiency and energy efficiency of perceptron have been significantly improved. With the continuous advancement of VLSI technology, the application of perceptrons in hardware accelerators, embedded systems and deep learning will be further expanded, driving more efficient neural network computing and real-time data processing.

4. Applications of VLSI in Artificial Neural Networks

Convolutional neural network, a branch of ANN, is a deep learning model used to process two-dimensional data such as images, videos, and speech. It performs well in image classification, object detection and semantic segmentation. However, CNNS have extremely high computational requirements, especially when dealing with high-resolution images and complex network architectures, involving a large number of convolution operations and matrix multiplications. Although the traditional architecture can process convolution operations in parallel, it has high power consumption. The VLSI architecture allows convolution operations to be assigned to multiple parallel computing units, which speeds up processing. By computing multiple convolution cores in parallel, hardware resources can be effectively utilized and the overall computing time can be reduced. In addition, in order to reduce external memory access, VLSI chips integrate on-chip cache, so that feature maps and convolution cores can be stored locally and accessed quickly. This significantly reduces the memory bandwidth bottleneck and improves the efficiency of data processing. The VLSI based CNN accelerator specifically accelerates the convolution calculation through hardware

optimization, which significantly improves the efficiency of the system. Literature studies have shown that CNN accelerators designed using VLSI can reduce energy consumption by a factor of 5-10 and compute latency by a factor of 2-5, resulting in a significant increase in efficiency compared to traditional GPU implementations. For example, for image classification tasks, a VLSI based CNN accelerator achieves processing speeds of up to 30 FPS while maintaining low power consumption.

5. Challenges in VLSI Implementation

In the process of mapping ANN to VLSI hardware, engineers face several technical challenges. These challenges are exacerbated by the increasing demand for computing power. There are potential problems in power management, computational resource optimization, model quantification, etc. ANN, especially deep neural networks, involve a large number of matrix multiplication and addition operations, which require a high density of computing resources. With the increase of the scale of neural networks, the hardware implemented by VLSI faces a huge power consumption problem. High power consumption will not only increase energy costs, but also cause the chip temperature to be too high, which affects the stability and life of the device. Engineers are trying to design computing units using low-power CMOS technology to reduce power consumption from circuit switching. In addition, ANN has a huge amount of computing, so it needs a huge amount of storage. ANN typically has millions or even billions of parameters. The hardware implemented by VLSI requires frequent access to weights and input data, which results in a significant increase in memory requirements and bandwidth requirements. Traditional external memory access creates latency, which affects overall performance. But engineers developed On-chip Cache technology. An On-chip Cache is a type of high-speed memory integrated into a processor or chip for quick access to frequently used data and instructions. Compared to external memory (such as DRAM), on-chip caching has faster access speed and lower latency, so it can significantly improve system performance. By storing key data inside the chip, it reduces the frequency of the processor's access to the external memory, thus speeding up the speed of data processing and reducing the bottleneck of memory access. However, the storage bottleneck is still a problem that has not been fully solved, especially when dealing with large-scale neural networks. Although the expansion of on-chip cache is a potential solution, the physical area of the chip is limited, and too large cache will not only increase the manufacturing cost, but also affect the heat dissipation performance of the chip. In addition, increasing the bandwidth of memory also faces bottlenecks in physical circuit design, especially at high frequencies where maintaining signal integrity becomes more difficult. Moreover, there is a natural conflict between designing low-power storage systems and improving storage performance. With the increase of storage bandwidth and capacity, power consumption inevitably rises, and how to maintain high performance while reducing energy consumption remains a difficult challenge to solve.

6. Conclusion

Through the discussion in this paper, VILSI technology plays a crucial role in the implementation of artificial neural networks. VLSI can not only significantly improve the processing efficiency of neural networks through parallel computing, low-power design and on-chip cache optimization strategies, but also effectively reduce the power consumption of hardware. In addition, the paper explores in depth the role of tree adders and accumulators in accelerating neural network operations, showing how VLSI architectures can further improve the speed of deep learning models by reducing the latency of addition and accumulation operations. However, while VLSI has made significant progress in optimizing ANN hardware designs, there are still many challenges in design as neural networks continue to grow in size and complexity. However, as the scale of neural networks continues to expand, problems such as storage bottlenecks and energy efficiency balance continue to plague designers and engineers. These problems which have not been completely solved need further technological innovation and research. In the future, with the emergence of new storage technology and more efficient architecture design, VLSI is expected to make greater progress in the field of ANN, and promote the hardware realization of neural networks to be more efficient and intelligent.

7. References

[1] Chen, Y., Emer, J., & Sze, V. Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. IEEE Micro, 2017, 37(1): 14-23.

[2] Wojcicki, T., & Iniewski, K. (Eds.). VLSI: Circuits for emerging applications. CRC Press, Taylor & Francis Group, 2015. Retrieved from Gallant, S. I. Perceptron-based learning algorithms. IEEE Transactions on Neural Networks, 1990, 1(2): 179-191.

[3] Zhu, X., Zhang, Y., Zhao, Z., & Zuo, J. Radio frequency sensing based environmental monitoring technology. In Fourth International Workshop on Pattern Recognition, SPIE, 2019, Vol. 11198: 187-191.

ISSN 2959-6157

[4] Vai, M. M. VLSI design. CRC Press, 2017.

[5] Ande, J. R. P. K., & Khair, M. A. High-Performance VLSI Architectures for Artificial Intelligence and Machine Learning Applications. International Journal of Reciprocal Symmetry and Theoretical Physics, 2019, 6(1): 20-30.

[6] Zhu, X., Zhao, Z., Wei, X., Wang, X., & Zuo, J. Action recognition method based on wavelet transform and neural network in wireless network. In Proceedings of the 2021 5th International Conference on Digital Signal Processing, 2021: 60-65.

[7] Wang, R., Zhu, J., Wang, S., Wang, T., Huang, J., & Zhu, X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. International Journal of Multimedia Information Retrieval, 2024, 13(4): 39.

[8] Amuru, D., Zahra, A., Vudumula, H. V., Cherupally, P. K., Gurram, S. R., Ahmad, A., & Abbas, Z. AI/ML algorithms and applications in VLSI design and technology. Integration, 2023, 93: 102048.

[9] Kaur, R., Asad, A., & Mohammadi, F. A Comprehensive Review of Processing-in-Memory Architectures for Deep Neural Networks. Computers, 2024, 13(7): 174.

[10] Jeyarohini, R., Sathya, R., Sellapaandi, S. P., Kavitha, P., Ramesh, D. R., & Mukherjee, A. An Examination of Machine Learning Techniques for Automating and Optimizing VLSI Design. In 2024 International Conference on Science Technology Engineering and Management (ICSTEM), IEEE, 2024: 1-6.