Artificial Intelligence(AI)chip technology review

Bingzhen Li
Naval aeronautical university
Yan Tai China
125395430@qq.com

Wenzhi Jiang
Naval aeronautical university
Yan Tai China

Jiaojiao Gu
Naval aeronautical university
Yan Tai China

***Abstract*—This paper takes the four most widely used Artificial Intelligence chips as the research object, analyzes its technology and architecture characteristics, summarizes its advantages and disadvantages, and points out the development trend of the next stage AI chip technology.**

Keywords-component; AI ship;GPU; FPGA; ASIC

I. INTRODUCTION

The algorithm, computing power and big data are the three elements that drive the rise of Artificial Intelligence. The three must be balanced and developed perfectly, and Artificial Intelligence can achieve good results. Computational power is the foundation of Artificial Intelligence. If computing power cannot keep pace with the development of algorithms and big data, Artificial Intelligence is also difficult to make breakthroughs. In recent years, due to the development of the big data industry, the amount of data has shown explosive growth, and the traditional computing architecture cannot support the massive parallel computing needs of deep learning. So the research community has carried out a new round of technology research and development on AI chips. Application research, foreign giants such as NVIDIA, Google, IBM and other international giants launched new products, domestic Baidu, Ali also have laid out the AI chip industry, and gave birth to such AI chip start-up company such as Cambrian[1].Chinese AI chip technology has achieved Great development.

II. AI CHIP BRIEF

The essence of Artificial Intelligence is an area derived from computers. Various types of chips play a role in the brain. The AI chip is one of the core technologies of the Artificial Intelligence era, which determines the infrastructure and development ecology of the platform. In a broad sense, any chip that can run Artificial Intelligence algorithms is called an AI chip, but in general, the AI chip refers to a chip that has been specially designed for Artificial Intelligence algorithms [2], and is currently used in the field of Artificial Intelligence. More chips are general-purpose chips (GPUs), semi-customized chips (FPGAs), fully-customized chips (ASICs) and brain-like chips. Among them, GPU, FPGA and ASIC all adopt the traditional von Neumann computer architecture, and brain-like chips use brain-like neural structures.

III. VARIOUS AI CHIP TECHNOLOGY FEATURES

At this stage, the GPU and the CPU are still the mainstream of the AI chip, and then with the continuous optimization of the algorithms for visual, voice and deep learning on the FPGA and ASIC chips, these two will gradually occupy more market shares, so as to achieve long-term coexistence with GPU. In the long run, artificial intelligence brain-like neural chip is the development path and direction. The characteristics of the four types of chips are described below.

A. GPU

The GPU (Graphics Processing Unit) image processor converts and displays the information to be displayed, provides scanning signals to the display, and controls the display of the computer. It is an important component for connecting the display and the computer motherboard. Compared to traditional CPU, GPU have a higher parallel architecture and are more efficient at processing graphics data and complex algorithms. Comparing the difference in structure between GPU and CPU, most of the CPU area is controller and register, while GPU has more ALU (ARITHMETIC LOGIC UNIT) for parallel processing of intensive data. The structure comparison is shown in Figure 1.

Figure 1. Comparison of CPU and GPU structure

Compared with single-core CPU, the running speed of programs on GPU system is often increased by tens or even thousands of times. For the processing of image data, every pixel on the image has a need to be processed, which is a

978-1-7281-5094-9/19/$31.00 ©2019 IEEE
DOI 10.1109/MLBDBI48998.2019.00028

considerable amount of data, so the field of image processing is the strongest in the field of computational acceleration. The GPU adopts single instruction and multi-data processing. It adopts a large number of computing units and an ultra-long pipeline. It mainly deals with the operation acceleration in the image field. It has strong versatility, high speed and high efficiency, and it is especially suitable for deep learning.

The original intention of the design is to cope with large-scale parallel computing in image processing. Therefore, when applied to deep learning algorithms, there are three limitations: First, the advantages of parallel computing cannot be fully utilized. Deep learning involves both training and inference. The GPU is very efficient in the training process, but in the inference process, the efficiency is general. Second, the hardware structure cannot be flexibly configured. The hardware structure of the GPU is relatively fixed, and the hardware structure cannot be flexibly configured like the FPGA. Third, the running algorithm is less energy efficient than the FPGA

B. FPGA

FPGA is a field programmable gate array, a semi-custom circuit in an application specific integrated circuit (ASIC). It is suitable for analyzing multiple instructions and single data streams. The advantages are low power consumption, high performance and programmable. Compared with CPU and GPU, it has obvious performance and energy consumption advantages, but it has high requirements for users. The user can define the connection between the gate and the memory by burning the FPGA configuration file. It not only solves the shortcomings of the flexibility of the custom circuit, but also overcomes the shortcomings of the limited number of gates of the original programmable device.

And FPGA can perform data parallel and task parallel computing at the same time, which has more obvious efficiency improvement when dealing with specific applications. The CPU needs multiple clock cycles for a particular operation. However, FPGA reorganizes the circuit through programming and directly generates the special circuit, which only consumes a small amount or even one clock cycle to complete the operation. In addition, many of the underlying hardware control techniques that are difficult to implement with general-purpose processors can be easily implemented using FPGAs. This feature leaves more room for the implementation and optimization of algorithms. Under the condition that the chip demand is not yet large, the deep learning algorithm is not stable, and iterative improvement is needed, it is one of the best choices to realize semi-custom Artificial Intelligence chip by using FPGA reconfigurable features.

In terms of power consumption, the FPGA also have advantage in terms of architecture. In the traditional Feng structure, an execution unit (such as a CPU) executes instructions, and the instruction memory, decoder, operator, and branch jump processing logic are required to participate in the operation. The function of each logic unit of the FPGA is reprogrammed. When it is entered, it has been determined that no instruction is required and no memory is shared, which can greatly reduce the power consumption of the unit execution and improve the overall energy consumption ratio.

Although FPGAs are highly optimistic, they are not developed specifically for the application of deep learning algorithms. There are also many limitations: First, the basic unit has limited computing power. Although there are a large number of very fine-grained basic units inside the FPGA, the computing power is much lower than that of the ALU modules in the CPU and GPU. Second, the proportion of computing resources is relatively low. In order to achieve reconfigurable features, a large amount of resources in the FPGA are used for on-chip routing and wiring. Third, the speed and power consumption still have a big gap compared with ASIC. Fourth, FPGA's price is higher than others

C. ASIC

ASIC is a dedicated AI chip customized to meet specific application requirements. It can be optimized at the hardware level, with small size, low power consumption, high performance and low cost. With the development of Artificial Intelligence algorithms and application technologies, and the gradual maturity of the ASIC industry environment, the fully customized Artificial Intelligence chip ASIC has gradually demonstrated its own advantages, which is very suitable for Artificial Intelligence application scenarios.

First of all, the performance improvement of ASIC is very obvious. NVIDIA's Tesla V100, for example, can provide tensor computing of up to 125 teraflops for deep learning related model training and inference applications, with data processing speed 12 times that of its GPU series launched in 2014.Google's TPU3.0, which USES 8-bit low-precision calculations to save transistors, speeds up to 100 PFlops(1,000 teraflops per second) and improves hardware performance to the level of the current chip seven years after its development by Moore's law. [4]. Just as the CPU changed the huge computer of the year, ASIC chip will also dramatically change the face of today's AI hardware devices. The famous Alpha Go beat Shishi Li using about 170 graphics processing units (GPUs) and 1200 central processing units (CPUs). These devices require a computer room, high-power air conditioning, and multiple experts for system maintenance. If all the dedicated chips are used, only one space of a normal storage box is needed, and the power consumption is also greatly reduced.

Second, downstream demand promotes the specialization of Artificial Intelligence chips. From servers and computers to driverless cars, drones, and smart home appliances, devices that are at least dozens of times larger than smartphones need to introduce cognitive interaction capabilities and Artificial Intelligence computing capabilities. For reasons of real-time and data privacy, these applications cannot be completely dependent on the cloud, and must have local hardware infrastructure support, which will bring a huge demand for Artificial Intelligence chips. Therefore, after the deep learning algorithm is stable, the AI chip can be fully customized by the ASIC design method, so that the performance, power consumption and area indicators are optimized for the deep learning algorithm.

The shortcomings of ASIC are also very obvious. The design and manufacture of ASIC requires a lot of money, a long development cycle and engineering cycle. Once the customization is difficult to modify, etc.

D. Brain-like chip

The brain-like chip is a chip that simulates the human brain neural network and can simulate the human brain structure for functional sensing[5]. There are hundreds of billions of neurons in the human brain. Each neuron is connected to other neurons through synapses, forming a huge neuron loop, transmitting signals in a distributed and concurrent manner. The computational power is extremely strong, as shown in Figure 2.

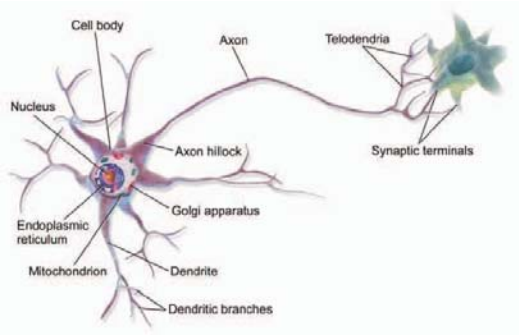

Figure 2. Human brain neuron structure

The research strategy of brain-like chips is to use hardware to imitate the synapses of the human brain, which is different from the traditional CPU structure, fully integrating the memory, CPU and communication components, using the digital processor as a neuron and the memory as a sudden touch. The processing of information is done entirely locally. As long as the neurons receive pulses from other neurons, they will simultaneously act to communicate with each other easily and quickly. Using the neural calculation method similar to the human brain, the energy consumption is low and the fault tolerance is strong. Compared with the traditional digital computer, the intelligence will be stronger, and the cognitive learning, automatic organization, and comprehensive processing of fuzzy information will also advance a big step. Therefore, brain-like chips are the future development direction of AI chips.

IV. THE DEVELOPMENT DIRECTION OF AI CHIP TECHNOLOGY

A. Technical defects of existing AI chips

At present, the core of mainstream AI chips is to achieve the speedup of the main convolution operation in CNN (convolutional neural network) by multipliers and accumulations. This generation of AI chips mainly has the following three aspects: First, the amount of data required for deep learning calculation is huge, and the memory bandwidth becomes the bottleneck of the entire system. Second, a large amount of memory access and MAC array computing, resulting in the overall power consumption of AI chips increased. Third, deep learning requires a lot of computing power. With the rapid development of deep learning algorithms, new algorithms are not well supported in accelerators that have been solidified. Therefore, the best way to improve computing power is to do hardware acceleration, which is to improve the computing power of AI chips.

B. The breakthrough direction of AI chips in the future

Therefore, it is foreseeable that the next generation of AI chips will have the following five trends.

First, more efficient convolution deconstruction / reuse.

Based on the standard SIMD, CNN can further reduce data communication on the bus due to its special multiplexing mechanism. The concept of reuse is particularly important in very large neural networks. How to reasonably decompose and map these super large convolutions to effective hardware has become a research direction.

Second, lower inference calculation / storage bit width.

One of the biggest evolutions of AI chips may be the rapid reduction of neural network parameters/calculation bit widths—from 32-bit floating point to 16-bit floating point/fixed point, 8-bit fixed point, and even 4-bit fixed point. In the field of theoretical computing, 2 or even 1 bit of parameter width has gradually entered the practice field.

Third, how to reduce the memory access delay.

When computing components are no longer the design bottleneck of neural network accelerators, how to reduce memory access latency will be the next research direction.

Fourth, a more sparse large-scale vector implementation.

Although the neural network is large, there are many cases where zero is input. At this time, the sparse calculation can reduce the useless energy efficiency, so as to reduce the useless power consumption.

Fifth, Computing and storage integration.

Process-in-memory technology, through the new non-volatile storage device, adds neural network computing function to the storage array, eliminating data moving operation, and realizes the neural network processing of computational storage integration, which significantly improves the power consumption performance.

V. CONCLUSION

In recent years, AI technology has made continuous breakthroughs. As an important physical foundation of AI technology, AI chips have great industrial value and strategic position. However, from the perspective of the general trend, it is still in the initial stage of the development of AI chips, and there is huge room for innovation in both scientific research and industrial applications. Only when the basic computing power reaches a certain height, and the synergy between the algorithm and big data can make Artificial Intelligence achieve a higher level of breakthrough.

REFERENCES

[1] JiKai, “A paper to understand tsinghua AI chip report, let you know at a glance on AI” ,November 2018 .

[2] Tsinghua University, “Artificial Intelligence Chip Research Report”. In Beijing, November 2018 .

[3] Yuan Chengyuan,“Semi-customized FPGA chip and fully customized ASIC chip” .Intelligence Car Science,December 2018.

[4] The playful mother of beans, “ The development is strong, the future can be expected. The past and future of AI chips”October 2018.

[5] Xu Weiming, “What kinds of artificial intelligence chips are there”, November 2018