ISSN: 2688-9323

NPU Development Overview

Zichao Wang

School of Computer Science and Technology, North China Electric Power University, Baoding 071003, China

Abstract

As a special embedded neural network chip mainly adopting ASIC techno- logy, NPU overcomes the inherent shortcomings of CPU and GPU in deep learning by means of hardware simulation neural network, and greatly improves the operation speed of deep learning chip. The Cambrian DianNao series of papers published in 2014 opened a precedent for the design of the dedicated artificial intelligence chip NPU architecture, which directly gave birth to the Cambrian series NPU, which to a certain extent led the Huawei Da Vinci architecture NPU, Ali "with light", Google TPU, etc. Although the main NPUs today are only focused on the field of inference chips, they have shaken the GPU's position in the field of artificial intelligence. The emergence of NPU represents the beginning of artificial intelligence chips in the direction of customization and specialization.

Keywords

NPU; neural network chip; DianNao; high performance; inference chip.

1. Introduction

Nowadays, the era of big data and intelligence is approaching, and the demand for deep learning computing is growing. The artificial intelligence chip market has begun to take shape and flourish. Artificial intelligence chips are mainly divided into training chips and reasoning chips, and application scenarios are mainly divided into cloud and terminal. Technically, the mainstream AI chips are divided into three categories: GPU, FPGA and ASIC chips. Nowadays, the AI training chips are mainly based on NVIDIA GPU, but in the field of reasoning chips, NPU customized new AI chips are emerging one after another. This paper will mainly discuss the development of NPU from three aspects: its concept, architecture and examples.

2. Concept

2.1. Definition

NPU (Neural-Network Processing Unit) is an embedded neural network processor, which adopts the architecture of "data-driven parallel computing" and is especially good at processing massive multimedia data such as video and image.

2.2. The Birth of NPU

For a long time, application requirements have been affecting the development direction of embedded technology. With the rise of deep learning neural network and the advent of artificial intelligence and big data era, CPU and GPU are gradually unable to meet the needs of deep learning. Facing the growing demand and the vast expected market, it is necessary to design an efficient intelligent processor specially used for neural network deep learning, so NPU came into being.

From a technical point of view, deep learning is actually a kind of multilayer large-scale artificial neural network. It imitates the biological neural network and is composed of several artificial

ISSN: 2688-9323

neuron nodes interconnected. Neurons are connected in pairs by synapses, and synapses record the weights of connections between neurons. Because the basic operation of deep learning is the processing of neurons and synapses, the storage and processing in neural networks are integrated, which are reflected by synaptic weights. However, in von Neumann structure, the storage and processing are separated, which are realized by memory and arithmetic unit respectively, and there are huge differences between them. When the existing classical computers based on von Neumann architecture (such as X86 processors and NVIDIA GPUs) are used to run neural network applications, they are inevitably restricted by the storage and processing separation structure, thus affecting the efficiency. Therefore, NPU, a specialized chip for artificial intelligence, has more necessity and demand for research and development.

2.3. The Difference between NPU, CPU, GPU

A CPU (central processing unit) is a central processing unit. It mainly includes an arithmetic unit (ALU) and a control unit (CU), and includes several registers, caches and buses for data, control and status communication among them. As the operation and control core of computer system, CPU is the final execution unit of information processing and program running. It is mainly responsible for multi-task management and scheduling. It has strong universality and is the core leading component of computer. Its computing ability is not strong, and it is better at logic control.

GPU (Graphics Processing Unit) is a kind of graphics processor, which can make up for the natural defects of CPU in computing power. Compared with the resources with less CPU and more cores, it uses a large number of computing units and an ultra-long pipeline, which is good at carrying out a large number of repeated calculations and speeding up the operation in the field of image processing. Its basic idea is parallel computing, that is, multiple processors solve the same problem together, and the solved problem is decomposed into several parts, and each part is calculated in parallel by an independent processor. However, his defects are obvious, that is, his coordination and management ability is weak, he can't work alone, and he needs CPU for control and scheduling. Although GPU is much faster than CPU in deep learning operation, it still has some problems, such as high-power consumption, easy overheating of chip, insufficient performance improvement and so on.

Compared with CPU, GPU is good at processing tasks and giving orders, while NPU is better at processing artificial intelligence tasks. NPU simulates human neurons and synapses at circuit level, and directly processes large-scale neurons and synapses with deep learning instruction set. One instruction completes the processing of a group of neurons. Compared with the von Neumann structure of CPU and GPU, NPU integrates storage and computation through synaptic weights, thus improving the operation efficiency. However, NPU also has some shortcomings. For example, it does not support the training of a large number of samples at present, and is relatively better at forecasting and reasoning.

2.4. Development Status of NPU

On June 20, 2016, the State Key Laboratory of Micro Digital Multimedia Chip Technology of Zhongxing announced in Beijing that it had successfully developed China's first embedded neural network processor (NPU) chip, becoming the world's first embedded video capture compression coding system on a chip with deep learning artificial intelligence, and named "Xingguang Intelligent No.1". As of May 2020, the manufacturers with NPU manufacturing capabilities include Cambrian Company (Cambricon-1H/1M, MLU100), Huawei (Shengteng), Ali (with light), Google (TPU), etc.

Compared with CPU, which has a relatively clear definition and a stable ecological chain, NPU is still in the ascendant stage, and it needs more places worth exploring and studying in both research and development and application. Compared with the dilemma of CPU and GPU

ISSN: 2688-9323

development and domestic catch-up, domestic manufacturers have undoubtedly taken the lead in NPU development, and the future can be expected.

3. Analysis of NPU Architecture

3.1. NPU Architecture

In 2014, Chen Tianshi's scientific research team of Chinese Academy of Sciences published DianNao series of papers, which immediately swept the architecture field and opened the precedent for designing special artificial intelligence chips. Later, CAMBRIAN Science and Technology under Chinese Academy of Sciences launched its first generation NPU CAMBRIAN 1A, which was used in Huawei Kirin 970 chip. Then Google introduced TPU architecture, Huawei introduced self-developed NPU based on Da Vinci architecture, and Ali introduced NPU with "with light" architecture. Subsequent NPU architecture is related to DianNao architecture. We mainly introduce DianNao architecture briefly.

3.2. DianNao

DianNao is the native architecture of CAMBRIAN NPU embedded processor, and it is the pioneering work of CAMBRIAN.

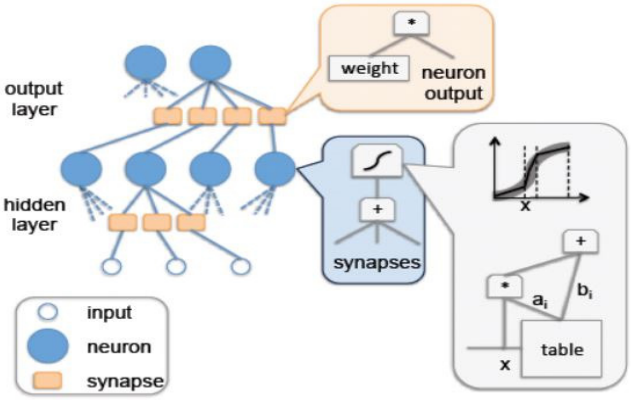

Figure 1. Complete hardware implementation of neural network

The above is the pattern diagram of neural network. Artificial intelligence algorithm based on neural network successfully simulates the structure of neurons in human brain. The neuron in the above figure represents a single neuron, synapse represents synapse of neuron, hidden layer is hidden layer in neural network, output layer is output layer, and input is input of neural network.

The following figure shows the internal structure of DianNao:

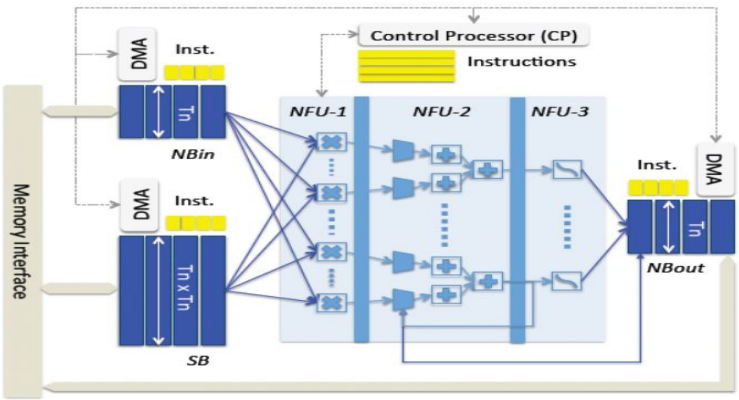

Figure 2. Accelerator

ISSN: 2688-9323

The blue area is the neural network structure simulated by hardware logic, which is called NFU (Neural Functional Units). From left to right, it is divided into three parts, NFU-1, NFU-2 and NFU-3. NFU-1 is a multiplication unit with 256(16*16) multipliers. NFU-2 is an addition tree. There are 16 addition trees, and each addition tree consists of 15 adders, which are arranged in the order of 8-4-2-1. NFU-3 is an activation unit with 16 activation units. Generally speaking, NFU divides resources into 16 parts, each part includes 16 multipliers of NFU-1, an addition tree (15 adders) of NFU-2 and an activation function operator of NFU-3. During operation, the multipliers in one part of resources run at the same time to output 16 results, which are sent to the addition tree. After the addition tree operation, a result is sent to the activation function, which determines whether it is activated or not.

In addition, there are three buffers, one for the input data (NBin), one for the operation weight (SB) and one for the result (NBout).

The performance of deep learning neural network chip based on DianNao architecture has been greatly improved, and its operation speed is far faster than GPU and CPU. The emergence of DianNao caused shock in the industry, and pioneered the special processor for deep learning neural network. Later, many different architectures emerged rapidly, the most famous of which was TPU of Google.

3.3. DaDianNao

Compared with the processor used by DianNao as an embedded terminal, DaDianNao is more suitable as a large-scale high-performance processor used by servers. In the design of DaDianNao, CAMBRIAN demands that the performance of DaDianNao should be improved by 16 times. Therefore, the scheme of expanding NFU resources by 16 times was initially adopted, but it was found that the wiring area was large and inefficient. Then, the multi-core parallel architecture was adopted, and the original NFU with expanded resources by 16 times was changed to 16 small NFUs. After reasonable wiring, the final area was reduced by 28.5% and the performance met the requirements.

3.4. PuDianNao

PuDianNao is an embedded processor scheme introduced by Cambrian in order to accelerate other important algorithms except deep learning in machine learning. PuDianNao internally implements seven commonly used machine learning algorithms: k-means, k-nearest neighbors, naive bayes, support vector machine, linear regression, and DNN. The structure of PuDianNao is as follows:

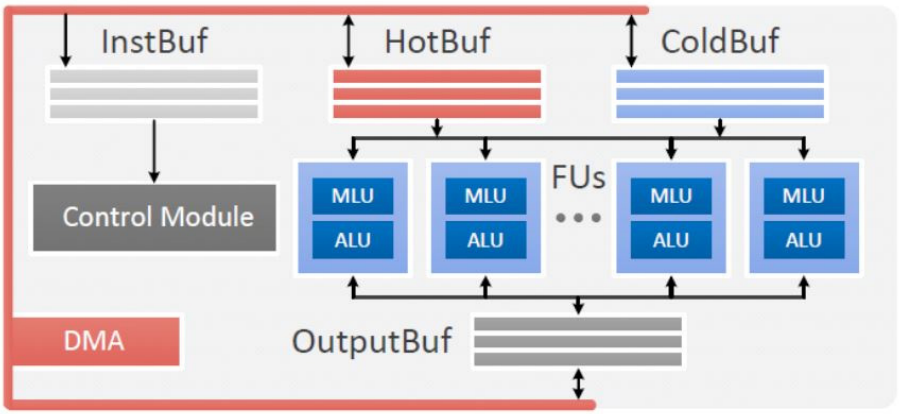

Figure 3. Accelerator structure of PuDianNao

ISSN: 2688-9323

The PuDianNao structure is similar to that of DianNao, including three buffers and multiple arithmetic units. The buffers include input data storage, weight storage and output data storage. The arithmetic unit consists of a plurality of FU (Function Unit) connected in parallel, and each FU contains an MLU (Machine Learning Unit) and an ALU. The MLU structure is as follows:

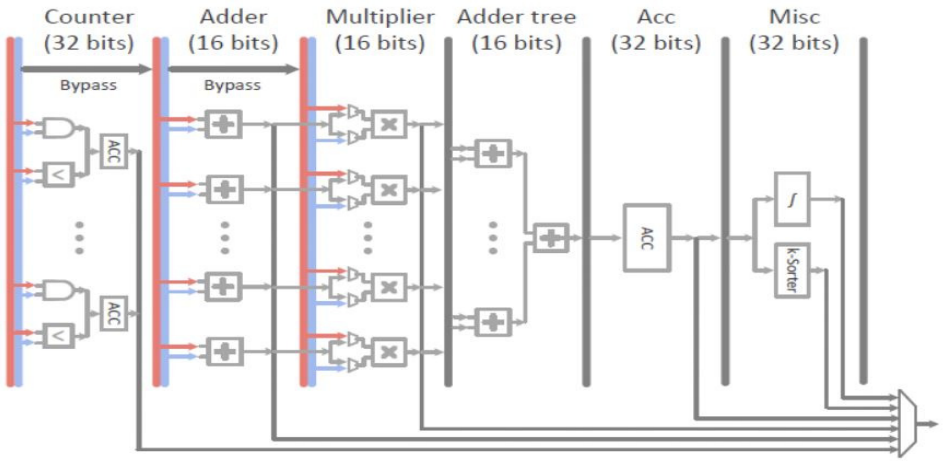

Figure 4. MLU hardware structure

The MLU structure is like NFU, except that two layers of logic Counter and Adder are added in front of NFU-1. Counter is used for accumulation, and the result is directly output to the next layer, which is mainly needed by naver bayes and classification tree. Adder is used in most machine learning algorithms, and the calculation results are directly output or used as the input of the next layer. Multiplier is equivalent to NFU-1,Adder tree is equivalent to NFU-2, ACC is used for accumulation, when computing resources are larger than hardware resources, calculated data can be directly stored in ACC for the next round of accumulation, saving the cost of reading and writing storage, Misc is equivalent to NFU-3.

The ALU contains an adder, multiplier, divider and a converter, which are used for some special MLU calculations in machine learning.

4. Examples of NPU Chips

4.1. Cambrian NPU

DianNao series of papers that swept the academic circle of architecture from 2014 to 2016 were published by the founder of Cambrian and his research team. In 2016, the company launched CAMBRIAN 1A, the first generation terminal intelligent processor IP product, which was the first commercial terminal intelligent processor IP product in the world. The integrated NPU used by Kirin 970, which claims to be the era of turning on mobile phone AI, is the CAMBRIAN 1A processor IP. In May, 2018, Cambrian officially released China's first cloud intelligent chip-Cambrian on MLU 100 chip, which indicates that Cambrian has become the first commercial company in China (and one of the few in the world) with both terminal and cloud intelligent processor products. The main products of CAMBRIAN Company are Cambricon 1H and 1M, which are the second and third generation IP architectures respectively.

4.2. Huawei NPU

Huawei began to use self-developed Da Vinci architecture to integrate NPU from Kirin 810 chip, which greatly improved the AI processing capability of the chip. Up to now, high-performance computing of mobile phone AI using Huawei NPU is still one of the selling points. In October 2018, Huawei also released the latest NPU chips, including Shengteng 910 for cloud training

ISSN: 2688-9323

and Shengteng 310 for terminal reasoning. Shengteng series NPU officially entered people's field of vision.

4.3. Alibaba Cloud "Han Guang" NPU

On September 25, 2019, Ali officially released the brand-new with light800 chip at the "2019 Yunqi Conference". "Han Guang 800" is a high-performance AI chip NPU using ASIC technology for cloud reasoning. The computing power of one "Han Guang 800" is equivalent to 10 GPUs, and the reasoning performance of "Han Guang 800" reaches 78563 IPS, and the energy efficiency ratio is 500 IPS/w. Compared with traditional GPU computing power, the cost performance is improved by 100%.

4.4. Google TPU

TPU is an ASIC-based computer neural network chip developed by Google in 2014, which is specially designed to accelerate the computing capability of deep neural networks. It is an acceleration operator focusing on neural networks and its development team draws lessons from the CAMBRIAN framework about DianNao series, so we also classify it as NPU in a broad sense here. The main innovations of TPU are the adoption of large-scale on-chip memory, integer operation with quantization technology (using 8-bit low-precision operation) and pulse array design.

5. Summary

The era of big data and intelligence is coming, and the application of deep learning training and reasoning is becoming more and more extensive. Artificial intelligence chips are ready to come out, and compared with the chip mode of CPU+GPU+FPGA, customized NPU mainly developed by ASIC technology is a symbol of the future. Although NPU can't compete with GPU in some specific fields, there is no doubt that NPU has shaken the position of NVIDIA GPU in the field of artificial intelligence. As long as it gives him enough time to grow, it will drive a new round of technological innovation in the industry.

References

[1] Tianshi Chen, Zidong Du,et al.DianNao[C],International Conference on Architectural Support for Programming Languages and Operating Systems. ACM, 2014:269-284.

[2] Luo T,Liu S,et al.DaDianNao[C],Ieee/acm International Symposium on Microarchitecture. IEEE, 2015:609-622.

[3] Liu D,Chen T, et al. PuDianNao[C], TwentiethInternational Conference on Architectural Support for Programming Languages andOperating Systems. ACM, 2015:369-381.

[4] Tian Ze, Some Thoughts on the Development of Embedded Technology [C], Micro-nano Electronics and Intelligent Manufacturing, Vol.2, No.1, 2020:1-4.

[5] evolove, AI chip: Cambrian NPU design analysis (DianNao): https:// blog. csdn. net/ evolone/ article/ details/80765094.

[6] Ali Dharma Institute: Algorithm and Architecture of with lightNPU: https:// blog. csdn. net/achuan 2015 /article/details/102680334.