International Journal of Psychosocial Rehabilitation, Vol.24, Issue 02, 2020
ISSN: 1475-7192

A history of development in brain chips in present and future

[1]Muhammad Talha,

Abstract:
Most high-speed calculations are based on big data set. However, constrained by Moore's Law, the number of devices cannot increase without limit. A suitable estimate with high accuracy and capacity is needed. Brain-like computation mimics the function of information analysing in human brains. Brain-like chips are hardware applying brain-like structure and analysis. Based on the neutron structure, brain-like chips can overcome the limitation of Von Neumann and improve both speed and complexity of calculation. The power consumption will decrease at the same time. In this essay, we will talk about the performances of several typical latest chips and discuss the difficulties and future development of brain-like chips.

Keywords: Brain-like computation, Brain-like chip, SNN

1. Introduction

Recent decades have witnessed a significant change in real-time computation. But with the limitation of Van Neumann's structure and Moore's Law, the calculation speed and physical size on the chip must be confined. The present solution is to adapt high-speed processors to accelerate calculation in MCU. Given Moore's Law, numbers of devices integrated on a single chip must meet its vertex. One possible solution is to find a more efficient structure to build up modern chips. Human brains have features of high efficiency, low power consumption, large capacity, and multiple-task-handling ability. Besides, they have strong learning ability, so chips modifying from this structure can be used in different environments.

Brain-like chips are based on human brains' biologic structure and brain-like computation copying information handling methods from mammals' brains. The spike carries information in organisms. The brain-like chip can not only do the series-parallel calculation but memorize lots of data. They have high calculation speed but low power consumption. The final aim is to make brain-like chips think and behave like a human.

Many research groups have published their results. However, these chips only work on the scale of a mature ape's mind. Both speed and accuracy, as well as learning ability, is so far from what we expected. Increasing the scale of calculation and number of "neuron unit" on a single chip are problems in this area. This passage will introduce components used in chip manufacture, biologic foundation, and brain-like calculation, compare performance chips produced by different companies. And we will also talk about biologic algorithms, learning rules, discuss difficulties, and future remedies.

2. Hardware and algorithm

This section will take about AI chips, typical brain-like devices, and several wide-used algorithms. We will also mention the latest brain-like chips around the world.

2.1 AI chips

AI chips can be divided into four types, which are GPU, FPGA, ASIC, and brain-like chip. GPUs are single instruction, but multiple data processed chips. They are mainly used on image processing and computational accelerating [1]. FPGAs are used in multi-instruction, single data flow analysis, widely used in behavior prediction [2], like cloud devices. ASICs are designed to meet users' needs; they are not changeable as long as they are manufactured. Brain-like chip mimic humans, particularly, is copying the human brain structure.

Type	Advantages	Disadvantages	Main usage	Typical company
GPU	General usage;	High power	Image	NVIDIA;

[1] *Department of Computer science, Superior University Lahore, Pakistan*

	High speed; High efficiency	consumption; Low efficiency in neutron network	processing	AMD
FPGA	Programmable Good performance Low power consumption	Normal performance in power consumption and general use	Behavior prediction (Cloud)	DeePhi Tech
ASIC	Task-targeted; Reliable; Controllable power	Long design and manufacture period; Not programmable	Movable devices	Google; Comparison; Horizon Robotics
Brain-like chip	Deficient power consumption; Short respond time;	Not reliable enough; In its budding period		IBM; West well Lab

2.2 Brain-like devices

Researchers should learn how human brains work to make computers or chips have 'thoughts' like mammals' brains. The whole action chain is getting input, analyzing data, learning and correcting, and then responding to a series of changes. It can use feedbacks to optimize the behavior and adjust the learning process.

a) MOSFET (CMOS, Complementary Metal Oxide Semiconductor). CMOS is a type of voltage-controlled devices. CMOS logic circuit can be applied to the voltage ranging from 5 volts to 15 volts, with lower noise and power dissipation. CMOS was widely used in chip manufacture in the past century. The technique and its application are very exhaustive. All aspects, including models, simulation tools, and manufacturing processes, are well developed. Also, CMOS can be used in VLSI design because of its highly integrated feature. In the traditional view, CMOSs are suitable devices to build up integrated circuits. Therefore, lots of engineers use them to construct brain-like chips at the beginning. However, the size of the CMOS cannot decrease without limitation. Existing lithography cannot make narrower gate width. 7nm process is considered a critical point in chip manufacture. And if we want to use CMOS to achieve the real function of our minds, the chip must be a huge one with billions of CMOS, the power consumption of whom must be an essential factor in the manufacture [3]. Although the whole circuit design and manufacture are complicated and burdensome, the human brain's architecture is not imitated. In conclusion, CMOS only can meet our demand for brain-like behavior [4].

b) Memory Devices. Memristors are kinds of nonlinear resistors, which are capable of memorizing the resistance under specific current flow. It will keep previous resistance even if the current is zero. Resistance will not change until inverse current flows. Memristors can store and represent data in a digital circuit, frequently used in RRAM. [5] In the brain-like circuit the resistance can be easily controlled, by changing the voltage applied to the mersisted. The chips are plastic. Besides, memristors have a smaller size than MOSFET, with lower power consumption [5]. Memristors can be integrated into three dimensions. Considering good knowledge of mersisters, many products are memory devices based. Generally, in memory devices, the SRAM is significant, and DRAM is volatile. The compatible memory is on the Nano-scale. Chips built with memory devices are programmed through data exchange. They need 1pA current to maintain states and 10pJ energy to rewrite. Therefore, if chips are built-in literary human brain size, the power consumption must be high, and memory is complex [6].

c) Artificial Synapse. Our brain works in a regular state with power consumption at around 20mW. And it can handle multiple tasks synchronously. Artificial synapse has lots of advantages in the latest human-made brain. Based on biological structure, modeling dendrite, soma, and axon, it can genuinely copy the neuron structure in a real physical brain. Therefore, the overall size of the chips could be smaller. The power-consuming is much less than extant chips. The connection between artificial synapse is strengthened during learning behaviour. In this emerging field, MIT Researchers has designed a computer chip working in analogy style, exchanging a gradient of signals. It changes weight like neurons in human brains, instead of carrying out computations based on binary [7].

Table 1. Comparison between several devices

Devices	Advantages	Disadvantages	Difficulties	Application
CMOS	Has well-developed standard	1.Complex	1.Method to decrease gate	True North

	manufacture process	2.Cannot imitate the structure of brains	width 2.Lithography (7nm)	(by IBM)
Memory Devices	1.Plastic 2. Easy to design 3.High efficiency 4.Low power consumption	1.SRAM is big 2.DRAM is volatile 3.Programmed through data change	Manufacture process and memory [2]	"Tianji" chip (by Tsinghua University)
Artificial Synapse	1.Memory with local computation 2.Low power consumption 3.Learn with spikes	Learning behaviour should take a longer time	Increasing size and complexity of the artificial brain	SiGe epitaxial memory for neuromorphic computing (By MIT)

Although CMOS is widely used in traditional chips, it is not suitable for brain-like structure because it is primarily constrained by manufacturing processes and physical limitations. Memristor is a better alternative and has many drawbacks compared to artificial synapses, like higher power consumption and lower efficiency. So artificial synapse will play an essential role in future research as a new-born structure.

2.3 Algorithm

The data of thirty college students' learning behaviours on the MOOC platform is shown as follows, representing any data object in class j.

Brain-like computation and brain-like chip design are meaningless without a bionic foundation. Brain-like calculation can be divided into three parts, including neurologic rules (especially knowledge of how the brain deal with much information simultaneously), hardware design, and brain-like learning method (or so-called software or platform). In reality, to make a chip or even computer think as human beings are commensurable as building up the biological structure of human brains and design a proper algorithm to model signals transferred in neutron. SNN (Spiking Neuron Networks) is the most similar biologic calculation method among extant algorithms. It is used to optimize the calculation both speed and accuracy in applications, companied with other algorithms. In this part, we will take about the bionic bases first, discuss the simplification model of simple minds, and comment on the results and performances.

a) Biological bases of SNN (Spiking Neuron Networks). Signals in mammals' brains are transmitted in neutron type of electronic signals. The ion concentration difference between the cell membrane's inner and outer membrane will lead to electronic voltage potential [8]. The peak of voltage potential represents the working state in organisms. In continuous-time, it will be shown as separate spikes.

b) Introduction of SNN. SNN is considered as the third generation of algorithms in artificial networks. In addition to neuronal and synaptic state, SNNs also incorporate the concept of time into their operating model. Neurons do not fire at each propagating circle; they work only when a membrane potential reaches a specific value. In spiking neural networks, the current often is considered as a different state of the artificial neuron. It will be fired when incoming spikes push the value higher and are quiescent when the value decay. SNN is different from other algorithms. Because of discontinuous work time, SNN has a quick response with much lower power consumption. The distinction between "fire" and "still" state is clear. But in contrast, the time period between every spike should be appropriately designed. One neutron should detect the signal, judge the country, and then dispatch information to the corresponding neighbor uni to imitate this structure. Artificial neuron network can be build based on knowledge of biological neuron structure [9].

c) Learning rule. Hebbian learning rule and STDP are two unsupervised learning methods. Hebbian learning is based on the Hebbian rule, which says when an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, the connection between them will be strengthened. This method can be used to explain combinational learning [10-11]. We can build up a relationship between the input layer and the output layer,

$$\Delta W_{ij} = \varepsilon x_i y_j$$

Where ε is the learning parameter, x_i and y_j is firing value of two neutrons, i is from 1 to n, and j is from 1 to m.

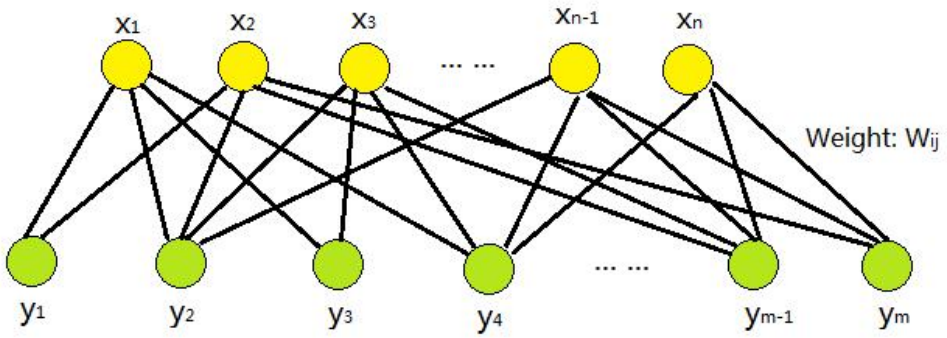

Figure 4. Simplified neutron network structure.

In the Hebbian learning method, take training data as input and use the model to get every weight, compare output and input, adjusting each weight by using the previous formula. If a new data set is written in the network, the value matched with pattern A will increase, and others will decrease [12]. STDP is another method derivate from the Hebbian learning rule. STDP (Spike-timing-dependent plasticity) can partly explain the activity-dependent development of the nervous system. It also describes the connection strength between neurons in the brain and makes the Hebbian learning rule accurate as tens of milliseconds. In the STDP process, the input will be strengthened if it occurs immediately before the output and weakens if it occurs immediately after the production. Thus the spike might be the reason for the post-synaptic neuron's excitation, which even likely to contribute a lot in the future. The connection between neutrons satisfy the relationship below,

$$\zeta(\Delta T) = \begin{cases} a^+ e^{-\Delta T/\tau^+} \\ -a^- e^{\Delta T/\tau^-} \end{cases}$$

Where a^+ and a^- are the extreme value of change in neutron, τ^+ and τ^- is the time constant? Although these two methods are used in many chips, the accuracy is low; its learning ability is limited. What is more, they need an extensive database to train the chip, and their unsupervised training is given some references.

3. Achievements and application

3.1 State of Art

After some decades of exploring, many teams have designed their human-like chips. We will list several typical digital and mixed-signal platforms in this section. True North constructed by IBM is a single chip (4.3cm2) with 4096 cores on it, the threshold voltage is near 0.53V in 28nm (Samsung) CMOS, the clock frequency is ~1kHz. Each core has 256 digital spiking neurons and 65kbit SRAM. It has configurable neuron parameters and state, AER-type packet-switched NoC, and router/core. [13] Loihi by Intel, Loihi is a 60mm2 chip in the 14nm CMOS process. It has 128 neuromorphic cores with three x86 cores, 0.75V supply, ~2pJ/spike-event. Each neuro core has 1024 spiking neuron primitives. Digital neuro cores support multi-compartment, dendritic, and several variants of STDP updates with reward modulation. The NOC mesh of it can scale to 4096 on-chip cores and 16384 chips through hierarchical addressing. [14] Each SpiNNaker board (from the University of Manchester) has 48 chips [15], 18 ARM cores. Each die (102mm2) has 128MB SDRAM 3D bonded. It consumes 1W per die (130nm CMOS) with the attached memory. It can code any neuron model and learning rules; a graph representation is used to build PyNN neural network. There are also some mixed-signal neuromorphic platforms, like Spikey chip by the University of Heidelberg and Dynamic Neuromorphic Asynch Processor (DYNAP). The comparison is listed in the form.

Name of chip	Features	Algorithm	Learning method	Virtues
TrueNorth (IBM)	CMOS digital circuit (5.4billion 28nm CMOS)	Convolutional networks	Offline training (transfer learning)	low power consumption ~70mW; corresponding simulation platform: Compass
Loihi (Intel)	CMOS, fully asynchronous implementation with barrier messaging	SNN (STDP)	On-chip learning	High learning efficiency; Low power;

				Core-to-core multicast
SpiNNaker (University of Manchester)	CMOS	SNN		Scalable and flexible architecture (graph representation builds PyNN)
Darwin (Zhejiang University)	CMOS, AER, each AER has time and neutron	ANN, SNN	Offline learning	Can avoid nondifferential feature of impulse
MLU100 (Cambricon)	MLUv01 structure and TSMC16nm process		Server: PHANERON	High speed; Low power consumption
Spikey (University of Heidelberg)	180nm CMOS		On-chip learning loop	Scaled to wafer-level BrainScale system
DYNAP (ETHZ/iniLabs)	180nm CMOS, external AER		On programmable chip connectivity	Small size; Low power consumption

TrueNorth can be used in many different circumstances, but the complexity of convolutional networks will increase dramatically with the increase of input data. TrueNorth and SpiNNaker are more users friendly because of the visible design platform. Darwin uses offline learning results of ANN into SNN. Therefore it will maintain the accuracy of data.

3.2 Applications

Nowadays, most chips are under test or just used on a small scale. But the idea of designing smart chips can be applied to many pragmatist fields, like writing recognizing, the intelligent assistant on the mobile phone, intelligent home automation, and intelligence robots. Brain-like chips will contribute a lot to logical analysis as well as intelligent recognizing. More wearable equipment will carry brain-like chips, aiming at being more human friendly and convenient.

4. Difficulties and further work

There are many existing problems through the past decades of developing of human-brain chips. The overall performances of extant chips are far from what researchers expected.

4.1 Biologic model and the structure

In electrical systems, the analog voltage carrying information becomes many separate spikes. The increasing and decaying period of voltage is eliminated; only the peak shows the state of the neutron. But the neutron is fired as long as the voltage potential reaches a specific voltage. The working time will be curtailed. With fragmentary information (details lost), rain-like chips will not respond to constructions correctly; the connection between different neurons will be chaotic. The learning process will be influenced. Besides, the size of the chip will be enormous if it acts like a real brain. The circuit design will be involved and demanding. With the increased size of the chip, the power consumption will be much higher. In this process, we can use analog signals to countervail the missing part of the movement. Shortly, engineers can put more effort into how to digitalize real spike in brains with more pulses in the artificial human brains. We can consider how to include rising time and reset time in the electrical system. Like dividing them into smaller timepieces or increasing numbers of spike taking information. However, that could be at the cost of a large number of artificial neutron and complex circuits. The basic structure of neurons is not enough for mimicking human brains. Researchers can think about the different working areas and build up neutron blocks [16]. On-chip design, we can have multiple logical layers, analysing data from lower to a higher order.

4.2 Calculation

The scale of calculation is small. The speed and accuracy are low. Compared to traditional chips, performances of brain-like chips are poor. To finish a single task, it will cost a lot of time. The simplified structure and model are limitations on mimicking the function of the simple mind. The process is just like a shrimp [17] or a mat, although it has attained resemble part as more than 80 billion neurons [18]. So it cannot handle the complex task—the time difference between galvanizing signal and response influences the calculation speed. Also, the nuance from the original signal can lead to preliminary results. The conmen method uses the traditional chip to carry an artificial neuron network to improve computation efficiency. As a transitional method, it can combine well-developed manufacturing processes of conventional chips and precise intelligence algorithms. Another threat is these chips based on artificial intelligence algorithms. However, AI is not an innovative idea, it was first mentioned in the 1980s, but few things new in this are these decadesa. So if AI calculation is up to its bound, the chip will not develop. Engineers should involve more types of accounting, data analysis methods in chip building.

4.3 Task handling methods and learning ability

Restricted by learning rules, brain-like chips have limit learning ability and learning areas. All extant fragments only can work on tasks sequentially. If it is converted into another environment, the training must be significantly different, and all the previous activity is in vain. But in human brains, parallel tasks are solved synchronously. Low efficiency in multiple task learning and handling will cost much more time and energy. Task block division will solve this problem [19]. Take smart robots, for example; having more sensors to detect changes in the environment is a promising solution. This feedbacks will provide details to optimize the training database. The training will be more targeted and efficient.

What is more, the primary general platform and build in instructions can be inserted on the chip. When used in a different atmosphere, they can be trained by the user through a series of interchangeable components. Feedbacks from sensors detect the change in the environment and provide more details in the training database. Training will be more targeted and efficient.

5. Conclusions

Brain-like chips are a new generation based on biological structures like human brains. They will contribute a lot to complex and large-scale computation in the future. Unsupervised learning rules on smart chips will help users to optimize interaction with the environment. But using brain science is a demanding interrelation field that needs lots of effort. Building a general platform to push multiple subjects learning will help chips attain good parallel task handling capacity. General platform and human-friendly compiler ensure wide use of the brain-like chip. Inspired by neutron blocks in human brains, chips can get used to a new environment with separation of the artificial synapse.

References

[1] John R Hu et. al., "Systematic co-optimization from chip design, process technology to systems for GPU AI chip", 2018 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA).

[2] Kesami Hagiwara et. al. "A two-state-pipeline CPU of SF_2 architecture implemented on FPGA and SoC for IoT, edge AI and robotic applications", 2018 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS).

[3] Jae-sun Seo et. al., "A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons", 2011 IEEE Custom Integrated Circuits Conference (CICC).

[4] Hideharu Amano et. al., "Panel discussions: 'Cool chips for the next decade'", 2017 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS).

[5] Hao Yu et. al., "Future brain-like computing with non-volatile memory device", 2016 China Semiconductor Technology International Conference (CSTIC).

[6] Cunhua Jia, "Development and future of memristor", p206-207, Digitization user(2013).

[7] Shinhyun Choi et. al., "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations", Nature Materials 17, p335-340(2018).

[8] C. Eliasmith et. al., "A large-scale model of the functioning brain", Science 338, p1202-1205 (2012).

[9] Luping Shi et. al., "Development of a Neuromorphic Computing System", IEEE IEDM, 978-1-4673-9894, p4.3.1-4.3.4 (2015).

[10] Yang Zeng-fang, Tang He-wen, "SNN Neighbor and SNN Density-based co-location pattern discovery", 2011 International Conference on E-Business and E-Government (ICEE).

[11] Aslihan Çavuş et. al., "Performance analysis of rule based automatic SNN algorithm on big data sets", 2018 26th Signal Processing and Communications Applications Conference (SIU).

[12] Pengfei Xie et. al., "Application of SNN-PID-based digital phase-locked loop to induction heating power supply", Proceedings of the 32nd Chinese Control Conference.

[13] Yinong Tang, "A history of brain-like chip in USA", Focus Technology (2015).

[14] M. Davies et. al., "Loihi: A neuromorphic manycore processor with on-chip learning", IEEE Micro (2018).

[15] Chong Chin Hui, Wang Ai Chie, "Chip on board (COB) versus board on chip (BOC) memory pachages", 2013 IEEE 15th Electronics Packaging Technology Conference (EPTC 2013).

[16] Eliasmith C, Anderson C H., "Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems", Cambridge, USA: The MIT Press, 2003.

[17] Kate D. Fischl et. al., "Path planning on the TrueNorth neurosynaptic system", 2017 IEEE International Symposium on Circuits and Systems (ISCAS).

[18] M. Davies et. al., "Loihi: A neuromorphic manycore processor with on-chip learning", IEEE Micro (2018).

[19] Vehi Calayir, Larry Pileggi, "Fully-digital oscillatory associative memories enabled by non-volatile logic", the 2013 International Joint Conference on Neural Networks (IJCNN).