Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/43/20230828

A review on neuromorphic computing circuits

Chen Jin

Department of Electronic and Electrical Engineering, University College London, WC1E 6BT, the United Kingdom

ken_j_kim@outlook.com

Abstract. Computational power is measured in terms of the time required for resolutions. The faster, the better. Due to its physical implementation, the conventional computing device used currently in computers is reaching its limit in computational power, and an essential innovation is required in order to break this limit. Neuromorphic circuits, which is inspired by neurology and simulates human biological brain and have higher functional efficiency to consume less energy and perform highly complicated tasks, is introduced and developed. This paper explains the principles of neuromorphic computing, which is representing ions in the biological neural system with electron in the circuit and adopting capacitors and resistors as counterparts of cellular membranes of the neural cells and the ion channel respectively. This paper gives some examples of neuromorphic circuits developed by several different corporations and laboratories. Algorithms mentioned in a certain research is exemplified and explained. Comparison of the difference between conventional and neuromorphic circuits is given to emphasize the advantage of neuromorphic circuits over the conventional ones. Several possible applications in a range of fields are also provided to depict the future prospect of this technology, including artificial intelligence, statistical calculation and information analysis. The conclusion is that the neuromorphic computers will replace the conventional Von Neumann computers, boosting the further development in computing power, breaking its limit.

Keywords: Electronics, Computer Sciences, Neuromorphic Computing.

1. Introduction

Human brains have been researched for several centuries. Ever since the invention of the very first computing machines, the attempt to simulate biological brain has been continuing. Despite that, the principles of how biological neural system and electronical computer are distinctly different. Computers function precisely and fast yet brains are more specified in solving new, complex, ambiguous and practical problems. Engineers take inspiration from neurons, which are the basic component of biological brains, and apply their working principles on electronic circuits. This is known as neuromorphic computing.

Different from an artificial neural network, which is achieved by pre-written programs that mimic human brains, neuromorphic computing requires physical simulation that is realized by constructing integrated circuits with artificial neurons.

According to Moore's law, the number of transistors on a microchip doubles every 18 months. [1] This has kept to be true for about half century due to that the size and price of CMOS has been decreasing and that the efficiency and speed has been increasing. Despite that, the production cost has also risen exponentially with the number of transistors, preventing the corporations gain profit before the next generation of product is introduced.

The advancement of microchip technology is driven by its widespread use in traditional computing applications. This is also accessible to neuromorphic systems. This phenomenon results in the enhancement of the size and functionalities of neuromorphic devices, however there is still a considerable distance to achieve a level of resemblance to biological systems. Consequently, the advancement of significantly large-scale neuromorphic systems is explored in this article.

2. Neuromorphic Computing Circuits

2.1. Principles of Neuromorphic Computing

Neuromorphic computing, as its name suggests, takes its inspiration from neurons in a biological brain. In 1952 Hodgkin and Huxley studied about biological neurons and derived a mathematical model of its mechanisms which produces 'spikes'.[2] The full biological neural system is shockingly complexed, but usually the mathematical models used will ignore some of its detail in order to build a physical projection which is a multiple-input-single-output system. Its input-output relationship can be described by several different mathematical functions.

Most of the process in the brain took place in the synapse instead of the neuron cell themselves. Synapses are the connections between neurons, through which they transmit the informatic signals. They are also simplified into several mathematical formulations. The models employed in this study effectively reflect the phenomenon of synaptic plasticity, wherein the synaptic effectiveness is dynamically adjusted to enable the network to learn the statistical properties of the inputs. Additionally, these models also incorporate structural plasticity, which involves the reorganisation of neurons to facilitate the storage of more enduring memories.

There is a large number of neurons and synapses in a mammalian brain. A human brain comprises 85 billion neurons connected by 10^{15} synapses.

In a synapse, signals are transmitted by passing ions through the membranes of neuron cells with some ionic channels connecting the two sides on them. In an electronic model, a capacitor is used to represent the membrane and a resistor connected parallel with it corresponds to the ionic channels. The ions are represented by the electrons.

The biological brain adopts digital technique for short distance signaling and analog chemical technique for long distance signaling. This approach differs from all-digital design of conventional general-purpose computers. In biology brains the logics are stochastic in order to achieve high efficiency. [3][5][6]

2.2. Examples of Neuromorphic Circuits

The IBM TrueNorth is developed using distributed digital neural models, which are anticipated to have real-time cognitive applications.

The Stanford Neurogrid employs sub-threshold analogue digital circuits that operate in real-time.

The Heidelberg BrainScaleS system utilises wafer-scale above threshold analogue brain circuits that operate at a speed 10,000 times faster than real-time biological processes. Its primary objective is to gain insights into biological systems and facilitate long-term learning.

The Manchester SpiNNaker machine is a computational system consisting of many cores that operates in real-time. It is designed to execute neural and synapse models by utilising software on compact embedded processors. Additionally, its purpose is to simulate biological nerve systems. The user's text does not contain any information to rewrite. [3]

2.3. Current Researches

The Institute for Neuroinformatics (INI) is renowned for its contributions to the field of neuromorphic computing. Notable areas of research at INI encompass the development of neuromorphic vision sensors, silicon cochlea, and medium-scale neuromorphic processors, such as the Reconfigurable On-Line

Learning Spiking (ROLLS) and cxQuad chips. The utilisation of subthreshold analogue circuits in the design of these devices is employed for the purpose of implementing spiking deep neural networks in display applications. Additionally, the circuit board is comprised of nine cxQuad chips and one ROLLS chip, which are visibly presented. The hierarchical convolution network is composed of cxQuad chips, each containing 1024 neurons and 65536 digital synapses. The classification layer is comprised of the ROLLS chip, which consists of 256 neurons and 128k analogue synapses. The technology in question exhibits reduced latency and power consumption compared to a conventional deep network operating on a large-scale digital cluster machine.

Furthermore, numerous universities are engaged in research within this field. For instance, UCSD has undertaken studies involving the creation of 65536-neuron two-compartment integrate-and-fire transreceiver modules. These modules incorporate spike-driven continuous time analogue membrane dynamics and are interconnected through a Hierarchical Address Event Representation (HiAER) communications fabric. Additionally, other universities have explored the integration of analogue neuromorphic circuit.

3. Algorithms and Applications

In order to facilitate the operation of a neuromorphic computer, it is necessary to develop a spiking neural network (SNN) that may be inputted into the computer. Spiking neural networks (SNNs) have computational characteristics that are reminiscent of biological neural systems. The majority of neuromorphic computers exhibit time-dependent characteristics in their implementation of neurons and synapses inside Spiking Neural Networks (SNNs). For instance, the phenomenon of spiking neurons involves the release of charges in accordance with a specific time constant. Additionally, both neurons and synapses exhibit a corresponding time delay.

The implementation of algorithms for neuromorphic systems often necessitates the establishment of a Spiking Neural Network (SNN) that is tailored to a certain purpose. The algorithmic methodologies employed in neuromorphic systems can be classified into two distinct categories: those that include training or learning a spiking neural network (SNN) to be integrated with a neuromorphic computer, and those that utilise non-machine learning techniques to manually configure SNNs for specific tasks. The terms "training" and "learning algorithms" in this context pertain to methods that modify the parameters of a Spiking Neural Network (SNN) in order to address a specific problem.

3.1. Machine Learning Algorithms

The proposed approach is a variant of backpropagation that utilises spike-based computations. The efficacy of backpropagation and stochastic gradient methods has been demonstrated in the context of deep learning. However, due to the non-differentiable nature of threshold functions in spiking neurons, these methods cannot be directly applied to spiking neural networks (SNNs). Furthermore, the temporal processing aspect of Spiking Neural Networks (SNNs) presents additional challenges in terms of training and learning for these methodologies. Algorithms designed for deep learning applications must be modified to accommodate Spiking Neural Networks (SNNs), which may result in a decrease in the accuracy of the SNN when compared to a comparable artificial neural network.

Several techniques that are commonly employed in deep learning training involve the utilisation of a surrogate gradient and the implementation of a smoothed activation function for the purpose of computing error gradients and changing weights across successive layers. There are instances where the computation of the spike error gradient exhibits similarities to the classification performance achieved by state-of-the-art methods on the handwritten dataset provided by the Modified National Institute of Standards and Technology (MNIST). Efforts have been made to use the intrinsic temporal dimension in Spiking Neural Networks (SNNs) by applying training procedures that have been traditionally utilised for recurrent neural networks, albeit with some approximations. Techniques such as backpropagation via time and real-time current learning have been used to neuromorphic datasets, specifically the Spiking Heidelberg Digits (SHD) and the Spiking Speech Command (SSC) dataset, as evidenced by previous studies.

The process of creating a representation of a pre-trained deep neural network. Given the operational training mechanism of deep neural networks (DNNs), certain endeavours to suggest a neuromorphic resolution for a specific problem often commence by training a DNN and subsequently converting the network into a spiking neural network (SNN) for the purpose of inference. Several of these approaches have demonstrated state-of-the-art performance while potentially reducing energy consumption. This is achieved by employing aggregate computations over both multiply and aggregate computations in deep neural networks (DNNs) on commonly used datasets, including MNIST, Canadian Institute for Advanced Research (CIFAR)-10, and ImageNet. Many original conversion strategies employed weight normalisation or activation normalisation, or opted for average pooling instead of max pooling. Additional techniques have been employed to train deep neural networks (DNNs) in a constrained manner, aiming to gradually approximate the activation function of a spiking neuron through repeated processes. Stockl et al. (year) have proposed a novel mapping technique wherein Spiking Neural Networks (SNNs) utilise the Few Spikes neuron model (FS-neuron). This model enables the representation of complex activation functions with a maximum of two spikes. These models exhibit similarities to deep neural networks used for image classification tasks, but with a reduced number of time-steps per inference compared to previously demonstrated approaches. Certain applications demonstrated using neuromorphic hardware have implemented mapping techniques that were previously stated. Efficient performance of tasks such as keyword search, medical picture analysis, and object detection has been demonstrated on established platforms such as Loihi by Intel and TrueNorth by IBM.

It is worth mentioning that the process of training a conventional deep neural network (DNN) and subsequently transferring it to neuromorphic hardware leads to a decrease in accuracy. This reduction in accuracy may be attributed not only to the transition from DNNs to spiking neural networks (SNNs), but also to the inherent characteristics of the neuromorphic hardware. The utilisation of neuromorphic hardware systems using developing hardware devices like memristors can frequently result in reduced precision of the synaptic weight values they are capable of achieving, as well as potential cycle-to-cycle device volatility. When developing a mapping technique, it is imperative to take into account the potential impact of certain properties on the inference performance of a mapped network. Moreover, algorithms that employ deep learning techniques for training spiking neural networks (SNNs) sometimes fail to fully exploit the intrinsic computing capabilities of SNNs. Consequently, adopting such approaches restricts the potential of SNNs to the achievements already demonstrated by conventional artificial neural networks. To illustrate, the majority of gradient-descent style algorithms, including mapping methodologies, do not prioritise the temporal dimension of spiking neural network (SNN) processing.

3.2. Reservoir computing

Reservoir computing, alternatively referred to as liquid state machines, represents another frequently employed approach in several domains. The approach introduces a sparse recurrent spiking neural network (SNN) that is denoted as a function referred to as "liquid" or "reservoir". Typically, this fluid is characterised in an arbitrary manner, although it must possess two essential attributes: input separability, which necessitates distinct inputs yielding distinct outputs, and fading memory, which mandates that signals within the reservoir do not perpetually propagate but rather diminish over time. In addition to the non-drained liquid component, a reservoir computing methodology encompasses a readout mechanism, typically a linear regression model, which is trained to discern and interpret the output generated by the reservoir. One notable benefit of this approach is its inherent capability to operate without requiring any training of the Spiking Neural Network (SNN) component. Reservoir computing in spiking neural networks (SNNs) employs sparse and recurrent connections, incorporating synaptic delays, within networks of spiking neurons. This approach facilitates the transformation of inputs into a higher dimensional environment that is characterised by both temporal and spatial dimensions. Certain instances of spike-based reservoir computing have demonstrated their efficacy in the processing of time-varying data. Different iterations of this computational framework have progressed from basic reservoir networks utilised for bio-signal processing and prosthetic control purposes to incorporating hierarchical layers of liquid state machines, which are a specific type of reservoir network. These layers are interconnected and trained specifically for video and audio signal processing applications.

3.3. Evolutionary approaches

These methodologies have also been employed for the training or design of SNNs. In the context of this specific technique, a population is generated by creating a random assortment of potential solutions. Every individual inside the population undergoes an assessment process where they are issued a numerical value referred to as a score. This score is subsequently utilised in the selection and reproduction stages to generate a new population. In the present context, evolutionary methodologies can be employed to determine the optimal parameters of the Spiking Neural Network (SNN), including neuron thresholds and synaptic delays, as well as the network's structure, encompassing the number of neurons and their interconnections via synapses. These methods are considered advantageous as they do not necessitate the presence of differentiability in the activation functions, nor do they depend on any specific network configuration. The network's structure and parameters can also undergo evolution as a result of these mechanisms. One drawback of evolutionary approaches is their more slower convergence compared to alternative methods. These methodologies have predominantly been employed in control-oriented contexts, such as video game development and the implementation of autonomous robot navigation systems.

3.4. Plasticity

Researchers have revealed that the modification of synaptic length, which is influenced by the activity of the interconnected neurons, has been proposed as a potential mechanism for learning in various activities. Spike-timing-dependent plasticity (STDP) is a widely studied synaptic plasticity mechanism in the field of neuromorphic literature. It functions by modifying the synaptic weights based on the relative timing of spikes between pre- and post-synaptic neurons. Various mathematical formulations pertaining to this topic are exemplified using the MNIST, CIFAR-10, and ImageNet datasets. Shresta et al. introduced a modified version of the exponential spike-timing-dependent plasticity (STDP) rule that is compatible with hardware. However, it should be noted that the classification performance on the MNIST dataset was not as high as the top-performing results achieved by spiking neural networks (SNNs). The resemblance between STDP-style rules and some machine learning methodologies, such as clustering and Bayesian inference, has been demonstrated in previous studies. STDP has been demonstrated as a spike sorting mechanism within the brain, functioning as a clustering mechanism. The integration of spiking reservoirs and spike-timing-dependent plasticity (STDP) has been employed inside a neural network framework known as NeuCube. This technique has been successfully utilised for the analysis of electroencephalograms and functional magnetic resonance imaging signals. Specifically, applications such as sleep state detection and prosthetic controllers have benefited from the implementation of NeuCube.

Recurrent spiking neural networks (SNNs) that incorporate delays and synaptic plasticity represent a broader category of models suitable for simulating dynamical systems. One example is the utilisation of polychronization networks in various spatio-temporal classification problems. Winner-take-all models have been demonstrated to enhance the classification capacity of recurrent spiking neural networks (SNNs). In order to accommodate the temporal component of Spiking Neural Networks (SNNs), some learning algorithms have been developed with the objective of producing singular or multiple spikes at specific time intervals. These methods have found utility in various classification tasks. The majority of these algorithms also depend on the spike representation employed to encode the input signals, which include spike rates, latency, and neuron population.

Non-machine learning algorithms refer to a class of algorithms that do not rely on the principles and techniques of machine learning. These algorithms are designed to solve computational problems by following a A prevalent category of algorithms that have been adapted for neuromorphic implementations is derived from graph theory. Neuromorphic computers operate on a directed graph structure. Consequently, if a compatible graph is available, it may be seamlessly integrated into the neuromorphic system, allowing for the identification of its inherent features using spike raster analysis. Neuromorphic computing was utilised in conjunction with graph theory to analyse the propagation of the COVID-19 epidemic. [4][5][6]

4. Comparison

In contrast to traditional von Neumann computers, which consist of distinct central processing units (CPUs) and memory units where both data and instructions are stored, neuromorphic computers employ neurons and synapses to govern both processing and memory functions. with the context of neuromorphic computers, programme definition is contingent upon the configuration of the neural network and its associated parameters, whereas with von Neumann computers, programmes are comprised of explicit instructions. Furthermore, in von Neumann computers, information is encoded through the use of numerical values that are represented by binary numbers. On the other hand, neuromorphic computers take input in the form of spikes, wherein the timing, size, and shape of these spikes are employed to encode numerical information.

Compared to conventional computing, it exhibits higher energy efficiency. The conversion between binary data and spikes is a subject of ongoing research, as the precise methodology for achieving this transformation remains an active topic of investigation.

In addition, there are other notable distinctions.

The operation of neuromorphic computers is characterised by a high degree of parallelism, wherein all neurons and synapses can simultaneously perform their respective functions. However, as comparison to von Neumann computers operating in parallel, the computations performed by neurons and synapses are very straightforward.

The concept of collocated processing and memory is observed in neuromorphic hardware, where the distinction between neurons as processing units and synapses as memory units is not always clear-cut. In the majority of cases, both neurons and synapses exhibit characteristics of both processing and memory functions. This approach aids in mitigating the von Neumann bottleneck, a phenomenon that arises from the inherent separation between the processor and memory components, resulting in a limitation on the maximum achievable throughput. Furthermore, this particular collocation serves the purpose of mitigating data accesses from primary memory, as this action often consumes a substantial amount of energy in comparison to computational energy in traditional computing systems.

The concept of inherent scalability pertains to the ability to expand the capacity of neuromorphic chips by incorporating a greater number of neurons and synapses than currently feasible. Multiple physical neuromorphic devices have the potential to be integrated and function as a unified hardware system, hence enabling the formation of expanded networks. This achievement has been successfully demonstrated on multiple occasions, exemplified by the SpiNNaker and Loihi systems.

Event-driven computation is a computational paradigm wherein the processing of information by neurons and synapses occurs exclusively when there are spikes, resulting in temporally sparse activity. This approach enables great efficiency in computational operations.

Stochasticity refers to the presence of inherent unpredictability in neuromorphic computers, where the firing of neurons introduces a certain level of noise. This does not present in conventional computers. [4]

5. Conclusion

It is very probably that neuromorphic computers would replace the traditional computers after its technology is advanced enough that its cost will be reduced and mass production of neuromorphic chips is possible, boosting functionality of all types of electronic devices and leading science technology into a next new era.

In the future the computer systems may adopt the stochastic logic inspired by biological brains. For example, the robot sensor systems, where absolute accuracy is not achievable, and energy efficiency is necessary. The neural and synapses within brains of humans of different ages can be studied to find out the optimum state of the biological neural system, and inspire neuromorphic hardware of higher level.

References

[1] Moore G E 1965 Cramming more components onto integrated circuits Electronics 38 114–7

[2] Hodgkin A and Huxley A F 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve J. Physiol. 117 500–44

[3] Furber, S. (2016). Large-scale neuromorphic computing systems. Journal of Neural Engineering, 13(5), 051001. https://doi.org/10.1088/1741-2560/13/5/051001

[4] Schuman, C. D., Kulkarni, S. R., Parsa, M., Mitchell, J. P., Date, P., &amp; Kay, B. (2022). Opportunities for neuromorphic computing algorithms and applications. Nature Computational Science, 2(1), 10–19. https://doi.org/10.1038/s43588-021-00184-y

[5] Marković, D., Mizrahi, A., Querlioz, D., & Grollier, J. (2020). Physics for neuromorphic computing. Nature Reviews Physics, 2(9), 499-510.

[6] Zendrikov, D., Solinas, S., & Indiveri, G. (2023). Brain-inspired methods for achieving robust computation in heterogeneous mixed-signal neuromorphic processing systems. Neuromorphic Computing and Engineering, 3(3), 034002.