Design and Analysis of Real Time Spiking Neural Network Decoder for Neuromorphic Chips

Chenyuan Zhao, Lingjia Liu, Yang Yi

The Bradley Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, U.S.
{zcyg542, ljliu, yangyi8}@vt.edu

ABSTRACT

Neuromorphic computing, which is based on non-traditional architectures that mimic bio-neurological process, could offer potentially disruptive capabilities and high energy-efficiency in real-time data monitoring and prediction, and resource allocation. The first step for de-signing neuromorphic chips is to explore effective design methodologies for real time and energy efficient spiking neural network (SNN) encoders and decoders. In this paper, a spike timing dependent plasticity principle (STDP) based decoder is designed and optimized. As shown in our experimental results, the proposed STDP based SNN decoder achieves good performance in information recovery with multiple scales.

KEYWORDS

Neuromorphic computing, encoder, decoder.

1 Introduction

As neuromorphic computing science has developed, analog computing has returned to the main stage in specific application fields. Brain inspired analog integrated circuits have made steady progress in the development of high-performance neurotrophic computing systems. Within such an analog neuromorphic computing system, the data interface, i.e. encoder and decoder, has played an important role since it determined the whole system's data through rate and efficiency, e.g. compact information format would consume fewer system resources and spend less time to process. The biological neuron have been modeled and implemented through either digital or analog techniques which served as the fundamental unit to build neural networks [1]. Most of these hardware implementations

of the neuromorphic computing are based on the rate encoding strategy [2], i.e. input signals are transformed into the firing frequency of a spike train. Another important information format, which has been proved to be existing in biological bodies, adopts the temporal-spatial status (phase) of spikes together with the inter-spike intervals (ISI) to carry the information [3]. Compared to the spike-based rate code, the spike-based phase code and ISI code have merits on higher information density and lower implementation cost [4]. Most importantly, both spike-based phase codes and ISI codes could be decoded back to the bi-stable information format by applying the spike time dependent plasticity (STDP) scheme [5].

In this paper, a STDP-based multi-scaled ISI decoder has been proposed and designed. The STDP-based ISI decoder could be used as the signal pre-processing interface between sensory information and spiking neural networks.

2 Neural Decoding

Not like traditional electronic systems which adopt eight digital level based signal or continue analog signal, spike train is the preferable information format in neuromorphic computing design. This is mainly because the spike is the dominating signal format in biological bodies. In most cases, the format of signals would determine the type of the processing structures and techniques in electronic circuits implementations. In [6], the design methodology and performance comparison of the neural encoding including rate code, latency code and ISI code have been discussed.

In the state of the art designs, integrate and fire (I&F) neuron is adopted to as the fundamental design unit. The firing frequency limitation of a general I&F neuron is mainly restricted by the membrane capacitor's value. For example, a typical 180nm complementary metal-oxide semiconductor (CMOS) based capacitor will only achieve as small as 30fF level. Such capacitance could be even smaller if parasitic capacitance is used to build I&F neuron. Therefore, the firing frequency of a typical I&F neuron could reach 10MHz or higher. However, the speed,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

ICONS '19, July 23–25, 2019, Knoxville, TN, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7680-8/19/07...\$15.00
<https://doi.org/10.1145/3354265.3354280>

or the throughput of the spiking neuromorphic computing system, is determined by the function blocks with slowest speed. Generally, the I&F neuron's firing frequency is depressed to fit the rest of the data link, i.e. the transmitting channel and decoding module. Between the aforementioned two modules, the decoding modules would be the bottle neck to the whole system since it needs not only transmitting spike signals but also transforming spike signals into digital or analog signals. Fig. 1 shows the working mechanism of the general decoding schemes.

Fig. 1 Spike processing schemes

As shown in Fig. 1, there are mainly four spike signal processing schemes. From top to bottom, the spike formats are digital rate spike code, analog spike rate code, latency spike code, and temporal spatial inter-spike interval spike code. The first processing method use the I&F neuron to transform sensory signal into rate spike train and adopt a digital counter to map the spike train into digital signal. The second and third processing methods also use I&F neuron to serve as the coding module. However, the spike firing frequency is much slower than the first one due to the processing speed limitation of STDP module. The fourth processing link's encoding and decoding modules are different from the first three schemes. I&F neuron cluster based ISI decoding and sampling based decoding schemes are adopted. Excepting the first scheme, the other three processing schemes need sample and hold (SH) module to serve as the post-processing links. The performance of the aforementioned four processing links is strongly depended on the decoding modules they use. For instance, counter for digital rate spike decoder, STDP for analog spike rate decoder and latency spike decoder, and ISI decoder (DE) for the temporal spatial inter-spike interval spike decoder. Therefore, by improving the decoder's performance, i.e. via improving its decoding scheme, the whole data link's performance, which including data rate and accuracy, would also increase.

Up to now, most state of the arts have adopted spike formats, i.e. spike rate code, spike latency code, to serve as information carrier in their designs [7, 8]. For the

spike-based ISI encoding technique, two types of the analog ISI encoder and one high sampling rate decoder are introduced in [6] [9]. However, among these implementations, there are many unsolved problems that need to be seriously considered, which are high error rate, low noise immunity, low speed limitation, limited scaling up capability, low reliability, and high power consumption. In the existing work [6-10], the error rate of the designs could reach to 0.1; the spike firing frequency is required to be less than 100KHz when STDP principle is adopted; and the scaling up ability is limited due to their design architectures [6-10].

3 ISI Decoding Methodology

In this paper, the STDP principle is applied to build an ISI spike decoder that could significantly improve the performance of the design in [9-12], i.e. the working frequency could be much higher than 100KHz. For an ISI spike processing module, one of the challenges is to design a robust and efficient decoder. There are two main steps in the encoding process: 1) convert the signal's amplitude into a latency spike, and 2) integrate latency spikes into an ISI spike train. Once the encoding process accomplishes, the numerical dimension (amplitude) of an input signal is transformed into a temporal dimension (time interval). Thereby, the purpose of the ISI decoding scheme is to transform the encoded temporal dimension based ISI spike train back to a numerical dimension based analog value, or a level-based signal.

In, a direct decoder design is discussed in which the key idea is to measure time intervals directly and transform these intervals into a level-based signals. However, this decoder adopts SH sub-circuits to detect the position of spikes, which occupy large active design areas. Most importantly, as the amount of spikes of an ISI code increases, more SH sub-circuits are needed. In other words, the reliability of the decoding operation has become a critical challenge when scaling up the information density of an ISI spike train. In order to overcome these issues, a decoding strategy, which includes the capability to handle any information density of an ISI spike train with a constant circuitry scaling factor, is needed. Inspired by the synapse updating technique, the STDP principle can be adopted to implement such a decoder for an ISI spike train.

One of the differences between the proposed ISI decoding and traditional STDP decoding scheme is the output scale, or amplitude level. Traditional STDP decoding could only generate bi-stable status which representing 1 and 0. However, the proposed ISI decoder could mapping ISI spike train into output with multiple

scales. By adopting the STDP principle, a novel decoder for an ISI spike train is proposed, as illustrated in Fig. 2.

Fig. 2 ISI decoder operation principle

As shown in Fig. 2, there are two channels that represent pre-spike and post-spike respectively. The time interval would have the same value as long as the encoding window is much smaller than the feature length of an input signal. By adopting the STDP principle, the output voltage could be expressed as

$$V_{out} = W(\Delta D), \quad (1)$$

in which the $W()$ is denoted as the learning function, e.g. the exponential function, which has been widely used in many designs [13–15]. For ISI spike code, the inter-spike intervals are changing continuously along with the input signal. Therefore, the STDP output would have several trends rather than bi-stable. The signal flow is illustrated in Fig. 3.

Fig. 3 ISI decoder signal flows

As shown in Fig. 3, by applying uniform reference spike train on the post-spike port and information carrier spike train on the pre-spike port, the final output from the ISI decoder would generate different voltages, i.e. V_1 , V_2 , V_3 . Under such configuration, there is no need to adopt additional SH module after the ISI decoder. Furthermore, the decoding speed increases significantly compared with the decoder in [9] [10]. This is mainly because the ISI spike train length is constant and much shorter than rate code spike train. Comparing with latency spike train, multiple spike based spike train has much higher error tolerance when spike missing is considered.

4 Circuit Design and Analysis

In this paper, a symmetric STDP-based circuit is designed to achieve the ISI decoding function, as illustrated in Fig. 4.

Fig. 4 STDP based ISI decoder circuit

As shown in Fig. 4, there are three major portions in this decoder design: the input module for the ISI spike train (constructed with transistors labeled in M_i), the input module for the reference spike train (constructed with transistors labeled in P_i), and the process module (constructed with transistors labeled in T_i).

In the input module for the ISI spike train, the input spike is applied on the gate terminal of M_1 and M_2 , which would be transformed into a spike train with the inverse voltage level. M_3 and M_4 would be used to transform the inverse spike train into a current to charge the C_1 capacitor. V_u is used to control the current intensity, and V_d is used to ensure M_5 working in the sub-threshold region. The input module for the reference spike train would have the same specification as the input module for an ISI spike train. These voltages across C_1 and C_2 , which are represented as V_1 and V_2 , would be applied on the comparator A_1 . Thereby, the output of A_1 could be expressed as:

$$V_3 = \begin{cases} V_{dd}, & V_1 - V_2 > 0 \\ 0.5V_{dd}, & V_1 - V_2 = 0. \\ 0, & V_1 - V_2 < 0 \end{cases} \quad (2)$$

Once the potential difference of V_1 and V_2 is compared by A_1 , T_1 and T_2 are used to charge C_r , while T_3 and T_4 are used to discharge C_r . Therefore, the output voltage of the decoder could be expressed as

$$V_{out} = \begin{cases} V_{ex}(1 - e^{-\frac{t_n}{\tau_n}}) \\ V_{ex}e^{-\frac{D_{ref}-t_n}{\tau_p}} \end{cases}, \quad (3)$$

where V_{ex} represents the transition voltage between the charging and discharging, D_{ref} is the inter-spike interval period of the reference ISI spike train, t_n is determined by V_3 , τ_n and τ_p could be expressed as

$$\begin{cases} \tau_n = C_r \left(\frac{W_{T4}}{L_{T4}} \right) I_{ON} e^{\frac{V_{p,q}}{nkT}} \\ \tau_p = C_r \left(\frac{W_{T1}}{L_{T1}} \right) I_{OP} e^{\frac{V_{n,q}}{nkT}} \end{cases} \quad (4)$$

where I_{ON} and I_{OP} are determined by the physical process of the NMOS and the PMOS respectively. By

regulating control voltages V_p and V_n , the same charging and discharging speeds could be achieved. In other words, the output voltage of the decoder is only determined by $V_1 - V_2$. In general, by adopting the proposed ISI decoder, the ISI spike train with different inter-spike intervals could be transformed into a level-based signal format.

5 Evaluation and Analysis

In this section, a hardware and software co-designed test bench is built to evaluate the comprehensive performance of the proposed ISI processor. The encoding and decoding accuracy is one of the most important characteristic. In the encoding design, I&F neuron is the key component to construct the proposed ISI encoder, thus, it is required to examine the power consumption of this neuron. The single neuron power consumption variation is illustrated in Fig. 5.

Fig. 5 Single neuron power consumption

In Fig. 5, it shows that as the input voltage increases, the power consumption would increase significantly. Therefore, it is better to constrain the input voltage within 0.4V to 0.55V. In this paper, the proposed ISI decoder is designed with the STDP principle. Unlike bi-stable STDP application, the proposed ISI decoder could achieve multiple scales by applying ISI spike code. The transient simulation results are illustrated in Fig. 6.

Fig. 6 Transient simulation results

It is clear that the proposed ISI decoder could achieve multiple scale decoding. In Fig. 6, 6-scale case is illustrated.

Acknowledgments

This material is based upon work funded by Air Force Research Lab (AFRL), under AFRL Grant No. FA8750-18-2-0009. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of AFRL.

REFERENCES

- [1] S. Furber, "Large-scale neuromorphic computing systems," *Journal of neural engineering*, vol. 13, no. 5, p. 051001, 2016.
- [2] G. Indiveri and T. K. Horiuchi, "Frontiers in neuromorphic engineering," *Frontiers in neuroscience*, vol. 5, p. 118, 2011.
- [3] D. S. Reich, F. Mechler, K. P. Purpura, and J. D. Victor, "Interspike intervals, receptive fields, and information encoding in primary visual cortex," *Journal of Neuroscience*, vol. 20, no. 5, pp. 1964-1974, 2000.
- [4] A. A. Grace and B. S. Bunney, "The control of firing pattern in nigral dopamine neurons: burst firing," *Journal of neuroscience*, vol. 4, no. 11, pp. 2877-2890, 1984.
- [5] T. Masquelier, R. Guyonneau, and S. J. Thorpe, "Competitive STDP-based spike pattern learning," *Neural computation*, vol. 21, no. 5, pp. 1259-1276, 2009.
- [6] C. Zhao et al., "Energy efficient spiking temporal encoder design for neuromorphic computing systems," *IEEE Transactions on Multi-Scale Computing Systems*, vol. 2, no. 4, pp. 265-276, 2016.
- [7] X. Guo, X. Qi, and J. G. Harris, "A time-to-first-spike CMOS image sensor," *IEEE Sensors Journal*, vol. 7, no. 8, pp. 1165-1175, 2007.
- [8] G. Indiveri, E. Chicca, and R. J. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE transactions on neural networks*, vol. 17, no. 1, 2006.
- [9] C. Zhao, et al., "Analog Spike-timing-dependent Resistive Crossbar Design for Brain Inspired Computing," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 8, no. 1, pp. 38 - 50, 2018.
- [10] C. Zhao, et al., "Making Neural Encoding Robust and Energy-Efficient: An Advanced Analog Temporal Encoder for Brain-Inspired Computing Systems," in *Proceedings of IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, 2016.
- [11] Y. Yi, et al., "FPGA based Spike Time Dependent Encoder and Reservoir Design in Neuromorphic Computing Processors," *Journal of Microprocessors and Microsystems: Embedded Hardware Design (Elsevier)*, vol. 46, Part B, pp. 175-183, 2016.
- [12] J. Li, et al., "A Deep Learning Based Approach for Analog Hardware Implementation of Delayed Feedback Reservoir Computing System," in *Proceedings of IEEE International Symposium on Quality Electronic Design (ISQED)*, 2018.
- [13] K. Bai, et al., "DFR: An Energy-efficient Analog Delay Feedback Reservoir Computing System for Brain-inspired Computing," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 14, no. 4, pp. 45-82, 2018.
- [14] Y. Dan and M.-m. Poo, "Spike timing-dependent plasticity of neural circuits," *Neuron*, vol. 44, no. 1, pp. 23-30, 2004.
- [15] K. Cameron, V. Boonsobhak, A. Murray, and D. Renshaw, "Spike timing dependent plasticity (STDP) can ameliorate process variations in neuromorphic VLSI," *IEEE Transactions on Neural Networks*, vol. 16, no. 6, pp. 1626-1637, 2005.