High Speed and Energy Efficient Deep Neural Network for Edge Computing

Kangjun Bai
Department of Electrical and
Computer Engineering
Virginia Polytechnic Institute
and State University
Blacksburg, Virginia, U.S.
kangjun@vt.edu

Shiya Liu
Department of Electrical and
Computer Engineering
Virginia Polytechnic Institute and
State University
Blacksburg, Virginia, U.S.
shiyal@vt.edu

Yang Yi
Department of Electrical and
Computer Engineering
Virginia Polytechnic Institute
and State University
Blacksburg, Virginia, U.S.
yangyi8@vt.edu

ABSTRACT

Edge computing enables data-stream acceleration with real-time data processing without latency, and allows for efficient data processing in that large amounts of data can be processed near the source with the ability to process data without ever putting it into a public cloud adds a useful layer of security for sensitive data. The edge computing-based architecture design and analysis play key impacts for the future Internet of Things (IoT) infrastructure development. In this work, we design a low power hybrid structured deep neural network (Hybrid-DNN), which employs memristive synapses working in a hierarchical information processing fashion and delay-based spiking neural network (SNN) modules as the readout layer, and provide a novel data layout method to allow the Hybrid DNN running a computationally intensive deep learning algorithm on limited resource edge devices. Motivated by the recent findings in neuromorphic computing and edge computing, we design a hybrid structured DNN combining both depth-in-space (spatial) and depth-in-time (temporal) deep learning architectures. Our Hybrid-DNN employs memristive synapses working in a hierarchical information processing fashion and delay-based spiking neural network modules as the readout layer.

CCS CONCEPTS

Edge computing

KEYWORDS

Edge computing, spiking neural network, deep learning

SEC '19, November 7–9, 2019, Arlington, VA, USA

ACM ISBN 978-1-4503-6733-2/19/11…$15.00
https://doi.org/10.1145/3318216.3363453.

1 Introduction

The continued success in the development of neuromorphic computing has immensely pushed today's artificial intelligence forward. Deep neural networks (DNNs), a brain-like machine learning architecture, rely on the intensive vector-matrix computation with extraordinary performance in data-extensive applications. Recently, the non-volatile memory crossbar array uniquely unveils its intrinsic vector-matrix computation with parallel computing capability in neural network designs [1].

Inspired by the neuroscience and proposed by Dr. Carver Mead in the 1980s, the neuromorphic computing has matured to provide intelligent systems that are capable of imitating neuro-biological processes through highly parallelized computing architectures [2]. As powerful as mammalian brains, the neuromorphic computing can efficiently solve complex tasks in applications on the pattern recognition and classification [3]. For instance, Loihi, the latest neuromorphic processor fabricated by Intel in 2017, recognizes a three-dimensional object from multiple viewing angles with a mere 1/1000 power consumption of the one used by a classic computer [4].

Deep neural networks (DNNs), the bio-inspired machine learning paradigm, have demonstrated extraordinary performance in data-extensive applications; for instance, the image recognition [5], the natural language processing [6], the autonomous driving [7], etc. The success in DNNs greatly promotes the development of hierarchical structured neural networks. Benefit by inherent recurrent connections with the depth-in-time deep learning characteristic, recurrent neural networks (RNNs) exhibit exceptional performance in processing temporal information. However, due to the status that neural information is propagated through numerous amount of synaptic weights with recurrent connections, training has become the most critical challenge for hierarchical RNN designs.

Edge computing enables data-stream acceleration with real-time data processing without latency, and allows for efficient data processing in that large amounts of data can be processed near the source with the ability to process data without ever putting it into a public cloud adds a useful layer of security for sensitive data. The edge computing-based architecture design and analysis play key impacts for the future Internet of Things (IoT) infrastructure development.

Motivated by the recent findings in neuromorphic computing and edge computing, we design a hybrid structured DNN combining both depth-in-space (spatial) and depth-in-time (temporal) deep learning architectures. Our Hybrid-DNN employs memristive synapses working in a hierarchical information processing fashion and delay-based spiking neural network modules as the readout layer. Major contributions of our work are summarized as followings:

- Our Hybrid-DNN combines both depth-in-space (spatial) and depth-in-time (temporal) deep learning characteristics, allowing such a system to carry out more sophisticated intelligent applications;
- With considerations of efficiency on the data processing, training mechanism is carried out through our fabricated delay-based SNN chip with the Mackey-Glass activation function, allowing previously learned knowledge to be transferred to next incoming inputs;
- We propose a novel data layout method to allow the Hybrid DNN running a computationally intensive deep learning algorithm on limited resource edge devices;
- Our prototype along with memristive synapses demonstrates its high computing parallelism and low hardware implementation cost with merely 1.05mW of on-chip power consumption.

2 Deep Delay Feedback Reservoir Design

The reservoir computing, an emerging RNN paradigm, is considered as a simplification of conventional RNNs that offer a unique training mechanism only at the readout stage. By only training output weights, the complexity of the training process is significantly reduced, resulting in higher computational efficiency. Despite that the reservoir computing has superior training performance, its hardware implementation is hindered by realizing the enormous amount of neurons, impeding such a powerful computing module to be deployed into portable devices.

Recently, a delay-based reservoir computing has been developed to reduce the hardware implementation cost; such delay-based system does not only have a resemblance to mammalian brains but also exhibits a near chaotic behavior, making the system an outstanding model for power-limited hierarchical RNN designs. Delay feedback reservoir (DFR) computing system [5] is a kind a recurrent neural network with part of weights are generated randomly. A DFR system [6], [7] consist of three parts which is mask layer, reservoir layer and readout layer. Mask layer serves as a mapping function that maps inputs to high dimensional data and Weights within the mask layer are untrained. Then, the mapped input is combined with output of reservoir layer from previous input to generate an input data for readout layer which computes the final output. Weights within readout layer are trained using gradient-based training method. Fig. 1 demonstrates the typical structure of DFR.

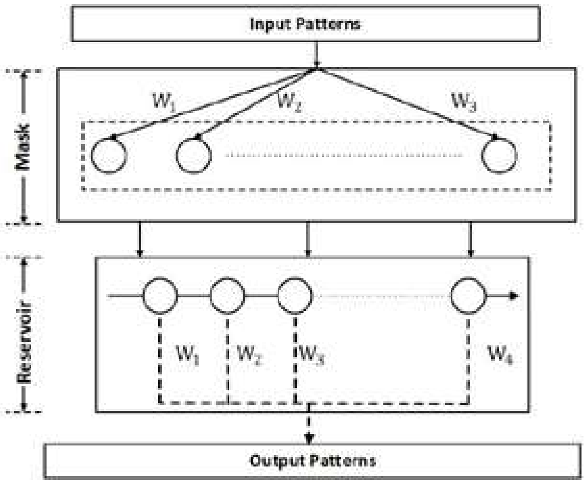

Fig. 1 Overview of the DFR system

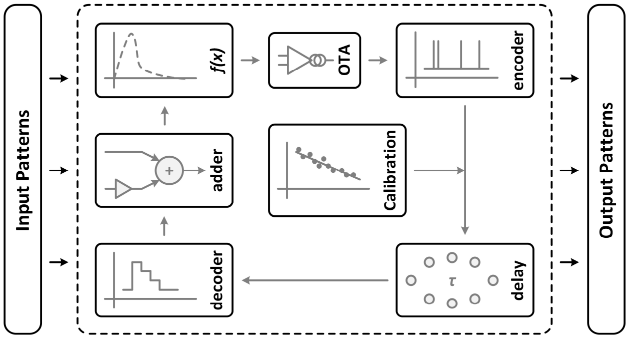

Fig. 2. General computing architecture of delay-based SNN

In this paper, we demonstrated a Hybrid-DNN computing paradigm with embedded memristive synapses and the spiking information processing technique., as shown in Fig. 2. In the endeavor to imitate how human beings process information, our Hybrid-DNN combines both spatial and temporal deep learning characteristic. The depth-in-space arises from the hierarchical structured neural network design, while the depth-in-time arises from inherent delay feedback connections within the delay-based SNN. We propose a novel data layout method to allow the Hybrid DNN running a computationally intensive deep learning algorithm on limited resource edge devices; Such a hybrid structured DNN is capable of handling data-intensive applications and establishing connections within the context of data.

3 Conclusion

Motivated by the recent findings in neuromorphic computing and edge computing, we design a hybrid structured DNN combining both depth-in-space (spatial) and depth-in-time (temporal) deep learning architectures. Our Hybrid-DNN employs memristive synapses working in a hierarchical information processing fashion and delay-based spiking neural network modules as the readout layer. Our Hybrid-DNN combines both depth-in-space (spatial) and depth-in-time (temporal) deep learning characteristics, allowing such a system to carry out more sophisticated intelligent applications; With considerations of efficiency on the data processing, training mechanism is carried out through our fabricated delay-based SNN chip with the Mackey-Glass activation function, allowing previously learned knowledge to be transferred to next incoming inputs. We propose a novel data layout method to allow the Hybrid DNN running a computationally intensive deep learning algorithm on limited resource edge devices;

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]] 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.

[4] 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.

[5] 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.

[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] 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.