Implementing Neural Networks on Memory-based Reconfigurable Logic Device (MRLD™)

Xihong ZHOU[1], Senling WANG[1], Yoshinobu HIGAMI[1], Hiroshi TAKAHASHI[1],
Masayuki SATO[2], Mitsunori KATSU[2] and Shoichi SEKIGUCHI[3]
[1]Department of Computer Science, Ehime University,
[2]TRL Co., Ltd., [3]TAIYO YUDEN CO., LTD.

With the rapid spread of artificial intelligence applications, neural networks (NNs) algorithm has achieved significant contributions at the machine learning domains including computer vision, speech recognition, and robotics. In order to achieve high performance and energy efficiency, hardware design for NNs such as implementing a NN on FPGA, ASIC design for NNs is gaining great attentions. MRLD™ is a new type of reconfigurable device constructed by SRAM array which has the advantages, including small delay, low power and low production cost. It is therefore a promising alternative device for NNs application. In this paper, we propose a novel network structure named MNN for implementing a neural network into MRLD™ device. We perform an experiment using the MNIST dataset to confirm the effectiveness of the proposed network structure (MNN).

1. Introduction

Over the past few years, Artificial neural networks (ANNs), usually simply called the Neural networks (NNs) become one of the most popular algorithms of artificial intelligence applications that has achieved significant contributions in machine learning domains including computer vision [1], speech recognition [2], and robotics [3]. The NNs consist of the input, hidden, and output layers [4]. Each of the hidden and output layers is composed of elementary computational units named neurons [5]. The layers are connected by neurons performing the multiply-accumulation and activating operation with parameters (weights, bias). In a practical intelligence application, NNs usually have multi-layers with millions of parameters which requires high-performance computing device. In addition, with the rapid spread of IoT (Internet of Things) technology in both the industrial and consumer fields, NNs are usually applied into various edge devices (e.g.: battery-powered mobile devices, robots, etc.) for real-time processing, computing device for NNs with high-performance and low power consumption is required. Since a general-purpose processor performing intensive multiply-accumulation operation sequentially usually be low efficiency, low speed and energy-consuming. Dedicated hardware design to achieve high performance and energy efficiency for NNs application is gaining increased attention.

GPUs (Graphics Processing Units) apply single-instruction-multiple-data with parallel processing that can significantly speed up the training process of complicate NNs [6-9], however, usually accomplished with huge energy cost (e.g.: NVIDIA A100 Tensor Core GPU, the thermal design power (TDP) is 400W [10]).

ASIC design for NNs is another key strategy for achieving high performance and energy efficiency for NNs application, such as Edge TPU, NVIDIA Xavier, and NovuTensor achieved good energy efficiency [11]. However, the extremely high development cost might obstruct the application of ASICs for IoT system.

Reconfigurable devices such as the FPGAs allow the user to reprogram the functionality and routing in field that provide a flexible and scalable platform for implementing the NNs application with high-performance and low power consuming [12]. MRLD™ (Memory-based Reconfigurable Logic Device) is a new type of reconfigurable device which is under development as an alternative to FPGA for the application of next-generation IoT/AI edge devices [13]. In contrast to FPGA which requires largely programmable interconnect resources to realize the programmability, MRLD™ is constructed only by general SRAMs array in a special internal connection structure that offers many advantages including the small delay, low production cost and energy efficiency (low power). It is therefore a promising alternative Edge AI device for NNs application.

In this paper, we propose an approach to implement a neural network into an MRLD™ device. By analyzing the basic operation of NN neurons in MRLD™, we propose a novel network structure named *MNN* (MRLD™-based Neural Network) to adapt the special structure of MRLD™. We also perform an experiment using the MNIST dataset to confirm the effectiveness of the proposed MNN.

The main contributions of this paper are as follows.

1) The execution principle of NN neurons in MRLD™ is analyzed.
2) A novel network structure named MNN is proposed.

The paper is organized as follows. Section 2 analyzes the basic operation of NN neurons in MRLD™. Section 3 addresses the proposed MNN (MRLD™-based Neural Network) for implementing a neural network into an MRLD™ device and describe the characteristics of MNN. Section 4 shows the experimental results for

*MRLD™ is the registered trademark of TAIYO YUDEN CO., LTD.

confirm the effectiveness of the proposed MNN by comparing the same size of traditional NN. Section 5 concludes the paper.

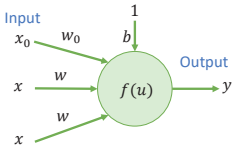
2. The basic operation of Neuron in MRLD™

In this section, we analyze the basic operation of NN neurons in MRLD™.

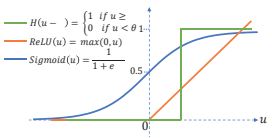
The single NN neuron is shown in Fig. 1 (a). The NN neuron's basic operation formula: $u = \sum_{i=0}^{N} w_i * x_i + b$, $y = f(u)$. The f is an activate function for u, the mainly used *Sigmoid*, *ReLU*, *Heaviside Step* (*Binary step*) activate functions is shown in Fig. 1 (b). As shown in Fig. 1 (c) the NN neuron's basic operation insinuates that the multiply-accumulate operation circuit and activate operation circuit and memory are necessary for NNs hardware design in traditional design method.

As shown in Fig 2 (a), an MRLD™ consists of multiple general-purpose memory cells (MLUT: Multiple Look-Up Tables) arranged in an array. The MLUT consists of two synchronous SRAMs (SRAM1, SRAM2) and two asynchronous SRAMs (SRAM3, SRAM4) as shown in Fig. 2 (b). According to the operating principle of MRLD™ [14][15], any computing function can be written into the MLUT in the form of a truth table. Therefore, in MRLD™, the NN neurons can be calculated in truth table form by binarizing inputs and outputs of NN which does not require constructing any logic circuits.

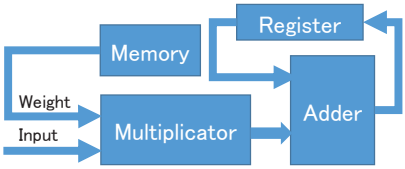
For illustration, we analyze NN neurons basic operation in MRLD™ with a size of $3 \times 3 \times 2$ ($l0, l1, l2$) NN. Fig. 3 (a) shows the NN, and parameters of each neuron. From the NN weights are extracted and form a mapping for x and output y shown in Fig. 3 (b). In MRLD™, as shown in Fig. 3 (c), the NN neuron's binarized inputs x^b as address-inputs of MLUT, through binarization function f^b calculate u to generates binarized activation y^b as data-outputs of MLUT. Therefore, as shown in Fig. 3 (d), the mapping of x^b and y^b can be calculated into a truth table stored in the MLUT to express the calculation operation of NN neurons.



(a) NN neuron

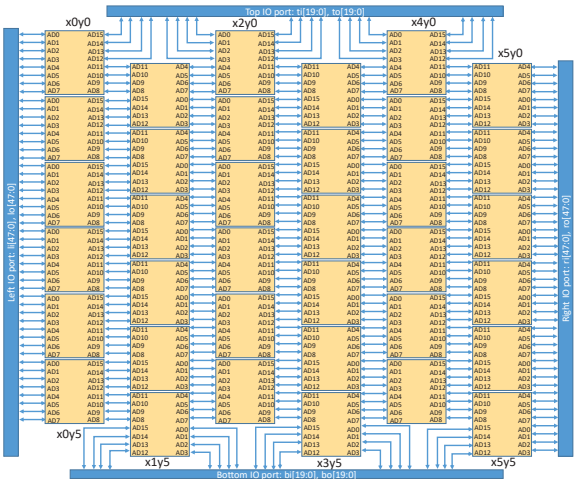


(b) Activation function $f(u)$

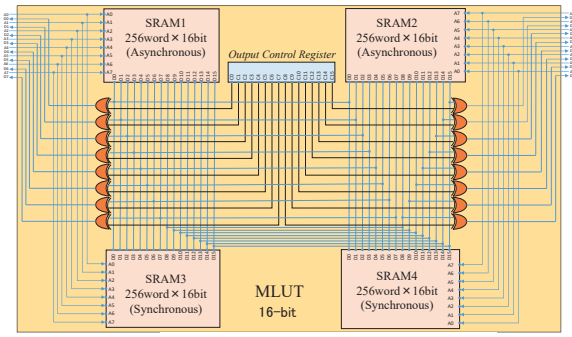


(c) Multiply-add and memory for neuron

Fig. 1 The NN neuron's basic operation



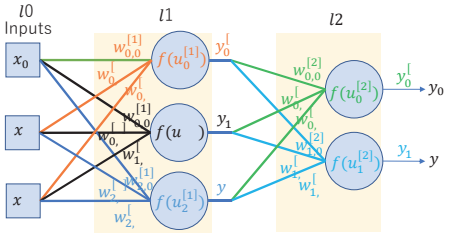
(a) MRLD™ (MLUTs array)



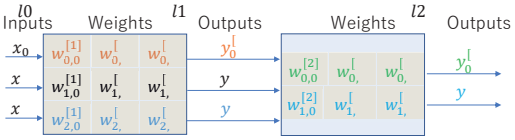
(b) Structure of a single MLUT

Fig. 2 MRLD™ structure

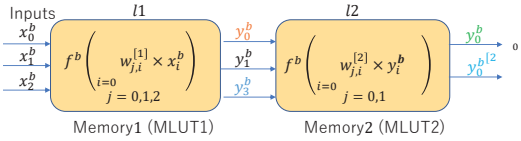
Note that we have not discussed the specific form of the binarization function for $x \to x^b$, $y \to y^b$, f^b and



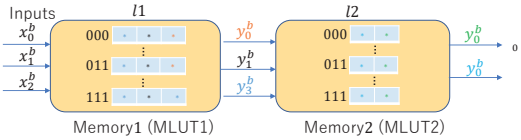
(a) A $3 \times 3 \times 2$ NN



(b) The parameters of NN neurons



(c) NN neurons operation in MULT



(d) NN neurons operation by the table form in MULT

Fig. 3 The NN neuron's basic operation in MRLD™

we will explore it in our future research. In this section, we analyzed the NN neuron's basic operation in MRLD™. The specific implementation analysis for NNs in the MRLD™ device, will be discussed in section 3.

3. MRLD™-based Neural Network: MNN

In this section, we explain and propose an *MNN* (MRLD™-based Neural Network) for implementing a neural network into an MRLD™ device. we also describe the characteristics of the MNN.

The traditional NN is a Fully Connected Neural Network (FNN), as shown in Fig. 4, where all neurons of each layer are fully connected with the preceding layer. For the MRLD™ structure, as shown in Fig. 5 (a), each MLUT is connected to a maximum of four MULTs, i.e. the left and right sides of each MLUT can be connected to two MLUTs respectively. Thus, the traditional fully connection way is not suitable with the constructing the NN into the MRLD™. We have to consider the MLUTs connection way in Fig. 5 (b), where a sparse NNs [16] in the unit of MLUT could be constructed in the MRLD™. In this paper, we propose a sparse neural network based on the MRLD™ structure named MNN (MRLD™-based Neural Network) is shown in Fig. 6 as the original structure.

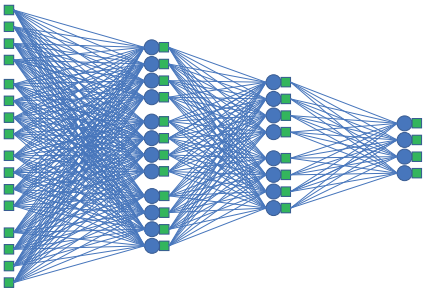


Fig. 4 The traditional Fully Connected Neural Network

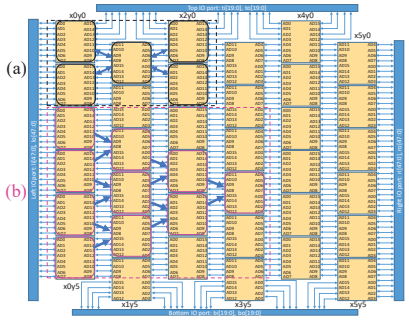


Fig. 5 The MLUT connection structure in MRLD™

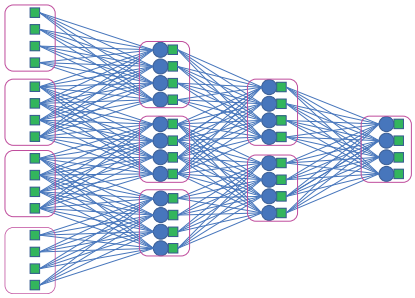


Fig. 6 The MNN (MRLD™-based Neural Network)

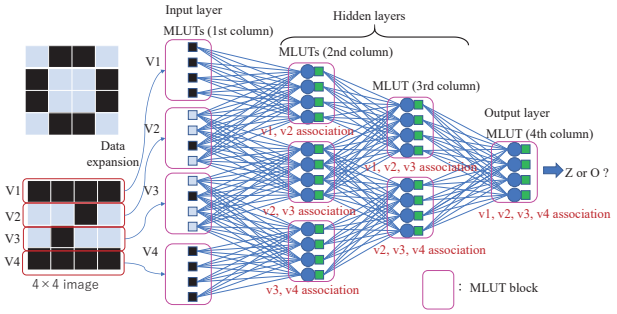
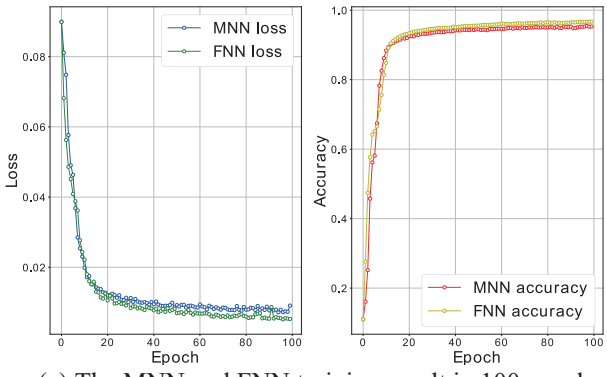


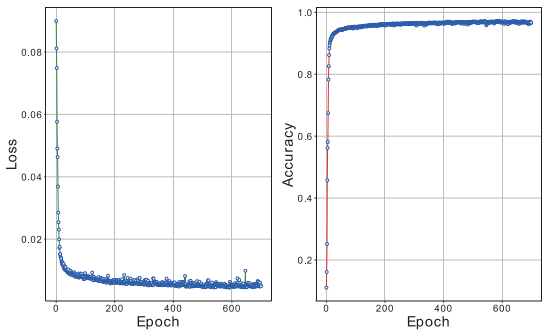
Fig. 7 The MNN feature

The MNN has inward gradual convergence and association characteristics, as shown in Fig. 7, In the input layer, data-input of each MLUT is independent with another MLUT, and the feature of these data will be converged and associated in the middle layer. For example, a 4×4 bit image of o and z is given respectively, and each 4bit is inputted in each MULT independently. Throughout the hidden layers, their feature is converged inwardly and gradually, and it is associated until the output layer, where all features are extracted and recognized.

In this section, we proposed and introduced an MNN for implementing the NNs into the MRLD™. In our future work, we will develop the method for image and data of any size that can be recognized in MNN in MRLD™. Before that, we need to confirm the effectiveness of MNN, see section 4.



(a) The MNN and FNN training result in 100 epochs



(b) The MNN training result in 700 epochs

Fig. 8 The MNN confirmation results

4. Experimental results

In this section, we show the experimental results to confirm the effectiveness of the proposed MNN by the training using the MNIST dataset.

In this experiment, for comparing with traditional NN, we first designed the same size of FNN and the MNN. we used the MNIST dataset (60,000 handwritten number training images and 10,000 test images.) to make training the MNN and the FNN, respectively. Fig. 8 (a) shows the training results in 100 epochs, the results show that the MNN is an effective neural network which can get well accuracy and loss as same as the FNN. Fig. 8 (b) shows the MNN has been 700 epochs trained, and it can obtain the training accuracy and testing accuracy up to 0.97 and 0.94, respectively.

5. Conclusions

In this paper, first, we introduced the MRLD™ device which is applied the next-generation Edge AI devices. For implementing the NNs into the MRLD™ devices, we first analyzed the NN neuron's operation principle in MRLD™. The NN neuron's operation can be calculated into truth table form storage in MLUT of MRLD™. In MRLD™, since each MLUT is connected to a maximum of four MULTs, it is difficult to construct the NN into the MRLD™ with the traditional fully connection way. Therefore, we have proposed a novel network structure MNN (MRLD™-based Neural Network) based on the MRLD™ structure. To confirm the effectiveness of the MNN, we performed a recognition training experiment using the MNIST dataset. The experimental results show the MNN is an effective neural network which can get well accuracy and loss for MNIST data recognition.

In our future work, we will explore the binarization method for the MNN and analyze design the method for image and data of any size that can be recognized in MNN in MRLD™.

Acknowledgment

This work was supported in part by KAKENHI (19K20234). In this research, the evaluation experiments for MRLD™ devices are supported by TAIYO YUDEN CO., LTD., and TRL Co., Ltd.

References

[1] K. He, X. Zhang, S. Ren and J. Sun: "Deep residual learning for image recognition", Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[2] T. Ochiai, S. Watanabe, S. Katagiri, T. Hori, J. Hershey: "Speaker Adaptation for Multichannel End-to-End Speech Recognition", Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 6707-6711, 2018.

[3] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, D. Quillen: "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection", Int. J. Robot. Res., vol. 37, no. 4-5, pp. 421-436, Apr. 2018.

[4] E. Grossi, M. Buscema: "Introduction to artificial neural networks", Eur. J. Gastroenterol. Hepatol., vol. 19, no. 12, pp. 1046-1054. Dec. 2007.

[5] W. S. McCulloch, W. H. Pitts: "A logical calculus of ideas immanent in nervous activity", Bull. Math. Biophys., vol. 5, no. 4, pp. 115-133. 1943.

[6] Z. Fan, F. Qiu, A. Kaufman, S. Yoakum-Stove: "GPU Cluster for High Performance Computing," Proc. ACM/IEEE Conf. on Supercomputing, Nov. 2004.

[7] E. Mizell, R. Biery: "Introduction to GPUs for Data Analytics Advances and Applications for Accelerated Computing", Published by O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.

[8] N. Singh, S. P. Panda: "Enhancing the Proficiency of Artificial Neural Network on Prediction with GPU", Int. Conf. on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Oct. 2019.

[9] Y. Kim, H. Choi, J. Lee, J. Kim, H. Jei, H. Roh: "Efficient Large-Scale Deep Learning Framework for Heterogeneous Multi-GPU Cluster", 2019 IEEE 4th Int. Workshops on Foundations and Applications of Self* Systems (FAS*W), Jun. 2019.

[10] NVIDIA: "NVIDIA A100 Tensor core GPU", NVIDIA, 2020.

[11] Y Hui, J Lien, X Lu: "Three-Dimensional Characterization on Edge AI Processors with Object Detection Workloads", Int. Conf. for High Performance Computing, Networking, Storage, and Analysis, Nov. 2019.

[12] K. Guo, S. Zeng, J. Yu, Y. Wang, H. Yang: "A Survey of FPGA-based Neural Network Inference Accelerators", J ACM Trans. Reconfigurable Technol. Syst., Vol. 12, No. 1, Article No.: 1, pp. 1-26, Apr. 2019.

[13] TAIYO YUDEN CO LTD: "Reconfigurable semiconductor device", Japan Patent JP2016208426A, Dec. 08, 2016.

[14] S. Wang, Y. Higami, H. Takahashi, M. Sato, M. Katsu and S. Sekiguchi: "Testing of Interconnect Defects in Memory Based Reconfigurable Logic Device (MRLD)", 2017 IEEE 26th Asian Test Symp. (ATS), Taipei, Nov. 2017, pp. 17-22.

[15] X. Zhou, S. Wang, Y. Higami, H. Takahashi: "Ring-Oscillator Implementation for Monitoring the Aging State of Memory-based Reconfigurable Logic Device (MRLD)", Int. Tech. Conf. on Circuits, Systems, Computers, and Communications (ITC-CSCC2020), Jul. 2020.

[16] S. Han, J. Pool, J. Tran, W. Dally: "Learning both weights and connections for efficient neural network", Advances in Neural Information Processing Systems 28 (NIPS 2015), Dec. 2015.