MNN: A Solution to Implement Neural Networks into a Memory-based Reconfigurable Logic Device (MRLD)

Xihong Zhou, Senling Wang,
Yoshinobu Higami and Hiroshi Takahashi
Dept. of Computer Science
Ehime University
Matsuyama, Japan
g863003a@mails.cc.ehime-u.ac.jp

Mitsunori Katsu
TRL Co., Ltd.
Tokyo, Japan
katsu@trl.jp

Shoichi Sekiguchi
TAIYO YUDEN CO., LTD.
Tokyo, Japan
s-sekiguchi@jty.yuden.co.jp

Abstract

MRLD™ is a new type of reconfigurable device constructed by general SRAM array (multiple-LUTs) which has the advantages including small delay, low power and low production cost. It is therefore a promising alternative device for Artificial Intelligence applications such as neural networks (NNs). However, implementing a traditional NNs with fully connected NNs is a hard task due to the special interconnection structure of SRAM array (the multiple look-up tables: MLUTs) in MRLD. In this paper, we suggest a LUT-based neuron model to realize neuron functions by writing truth table in SRAM array, and propose a novel neural network structure named MNN (MRLD-based Neural Network) to adapt the special connection structure of MLUTs for implementing a NNs application into MRLD. To evaluate the effectiveness of MNN, we perform the experiments by training MNN with the MNIST dataset. The experimental results show that the MNN can get almost the same accuracy and loss for MNIST data recognition compared to a fully connected NN.

Keywords: artificial intelligence, accelerator, reconfigurable device, LUT, MRLD.

1. Introduction

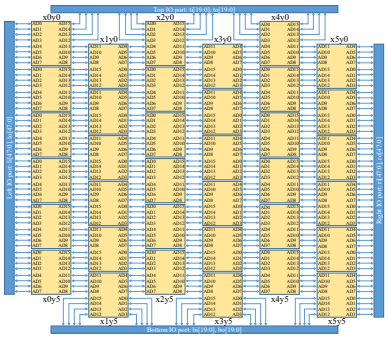
With the rapid spread of artificial intelligence applications, the neural networks (NNs) algorithm has achieved significant successes at the machine learning domains including computer vision [1], speech recognition [2], and robotics [3]. In a practical intelligence application, NNs usually consist of millions of parameters involving multiply-accumulation operations, which requires high-performance computing equipment. In addition, with the rapid spread of IoT (Internet of Things) technology in both the industrial and consumer fields, NNs are widely applied into various edge terminals, e.g.: battery-powered mobile devices, robots, electric vehicle etc.. In such systems, real-time processing, low power consumption and low cost are the main concerns with the computing device used for NNs [4]. In order to achieve high performance and energy efficiency for AI application, hardware design for NNs is gaining great attentions [5].

Over the past few years, the strategy of hardware design for NNs application can mainly be classified into three types: 1) Use GPUs (Graphics Processing Units) to accelerate NN training. 2) ASICs (Application Specific Integrated Circuits) design for NNs. 3) FPGA-based accelerators of NNs. The GPUs apply single-instruction-multiple-data in 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]) that is not suitable for edge device. The 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. Compared to ASIC design, reconfigurable devices such as FPGAs allow the user to reprogram the functionality and routing in field that can provide a flexible and scalable platform for implementing the NNs application with high-performance and low power consuming [12], however, the large area, delay and power issues due to the programmable interconnect resources prevent the use of FPGAs, and high cost is not friendly to the end user of edge devices.

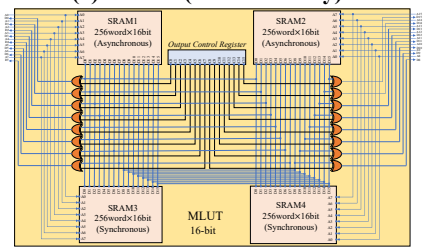
MRLD (Memory-based Reconfigurable Logic Device) is a new type of reconfigurable device which is under development as an alternative to FPGA for the application of next-generation IoT/AI edge devices [13]. In contrast to FPGA which requires largely programmable interconnect resources to realize the programmability, MRLD is constructed only by general SRAMs array (the multiple look-up tables: MLUTs) in a special internal connection structure that offers many advantages including the small delay, low production cost and energy efficiency (low power) compared to FPGA. In MRLD, functions (arithmetic logic, wiring logic) are expressed in the form of truth tables pre-stored in the SRAMs of MLUT. Since large amount of interconnect resources like in FPGA are not needed anymore, a large number of SRAMs can be integrated that provides a chance to implement large and complex functions in a single MRLD by truth tables, such like a LUT-based neuron activation function [14] and the LUT-based Neural Networks (L-NNs) instead of implementing an accelerator in FPGA (due to the limited memory size of LUT). Since the LUT-based neuron model [14] only operates memory, it thus would work much faster and low power than a traditional accelerator which has to perform the multiply-accumulation operations every cycle even though with acceleration circuits. Therefore, we believe that MRLD would be a promising alternative Edge AI device for NNs application.

On the other hands, due to the special structure of MRLD, implementing a neural network with fully-connected structure is an impossible task. There is an issue to implement an NNs application in MRLD, it needs a newly designed NN structure to adapt to the MRLD special structure.

In this paper, we suggest a LUT-based neuron model to realize neuron functions in truth table, and propose a novel neural



(a) MRLD (MLUTs array)



(b) Structure of a single MLUT

Fig. 1 MRLD structure

network structure named MNN (MRLD-based Neural Network) to adapt the special connection structure of MLUTs for implementing a neural network into MRLD. To confirm the LUT-based neuron model, we design a logic simulation experiment in an MRLD with 6×6 MLUTs array. The simulation results confirm the feasibility of LUT-based neuron function expression are the same as the results of the theoretical analysis. To evaluate the effectiveness of MNN, we also perform an experiment by training MNN with the MNIST dataset. The experimental results show that the MNN can get almost the same accuracy and loss for MNIST data recognition compared to a fully-connected neural network.

The main contributions of this paper are as follows.

1) A LUT-based neuron model is introduced.

2) A novel network structure named MNN is proposed.

The paper is organized as follows. Section 2 suggests a LUT-based neuron model. Section 3 proposes an MNN (MRLD-based Neural Network) for implementing the NNs application into an MRLD device and describes the characteristics and wiring connection way of MNN. Section 4 performs two experiments for confirming the LUT-based neuron model and evaluating the effectiveness of the proposed MNN, respectively. Section 5 concludes the paper.

2. LUT-based neuron model

In this section, we suggest a LUT-based neuron model to realize neuron functions in truth tables in MRLD.

Fig. 1 (a) shows the overall structure of a MRLD consists of 6×6 MLUTs (Multiple Look-Up Tables) array in mesh connection. Each MLUT consists of two synchronous SRAMs (SRAM1, SRAM2) and two asynchronous SRAMs (SRAM3, SRAM4), see Fig.1 (b). Between the MLUTs, address input lines and data output lines are bidirectionally interconnected in pairs (called AD pairs), and each MLUT is connected to four adjacent MLUTs by quarter of its AD pairs (except the outermost MLUTs). In such architecture, each SRAM works as a single look-up table (LUT), user can configure logics, wires function in the MLUT by writing

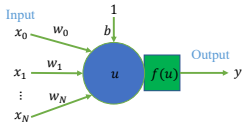
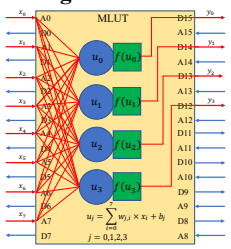
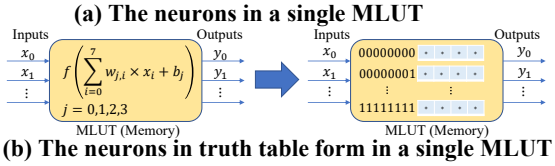


Fig. 2 NN neuron



(a) The neurons in a single MLUT



(b) The neurons in truth table form in a single MLUT

Fig. 3 LUT-based neuron model in a single MLUT

the corresponding truth tables in the SRAMs. According to the operating principle of MRLD [15][16] (our previous work for MRLD testing), any functions (including wiring logic) can be written into the MLUT in the form of a truth table that provides a new computing model for neuron activation functions in MRLD. In addition, a large number of MLUTs make it possible to implement a complete LUT-based NN into a single MRLD device.

Fig. 2. shows a basic neuron function expressed in formula: $u = \sum_{i=0}^{N} w_i \times x_i + b$, $y = f(u)$, where f is an activate function for u. To compute the value of y for u, a traditional approach has to perform multiply-accumulate operation and activate operation in many cycles, and requires large memory (buffer) to store the weights and input/output vectors.

The main idea of this study is that a neuron function can be expressed in a truth table in MRLD. Fig. 3 (a) shows an example to implement four neurons function in one MLUT. The correspondence between inputs x and outputs y of the neurons can be computed by pre-learning and formed in a truth table in the MLUT between the address-inputs and the data-outputs. Therefore, as shown in Fig. 3 (b), when calculate the output y for a given input pattern x, it only needs to access the memory and readout the prestored results of y. It is thus much faster and low power than a traditional accelerator which has to perform the multiply-accumulation operations every cycle even though with acceleration circuits.

Note that, here we are discussing the binarization form for x, y. For the specific binarization method for x and y, we will explore it in our future research.

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 and introduce the implementation way of the MNN neurons in MRLD.

3.1 A sparse neural network: MNN

A Fully-connected Neural Network (FNN) cannot be directly constructed into MRLD. As shown in Fig. 5, all neurons of each

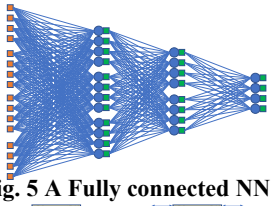


Fig. 5 A Fully connected NN

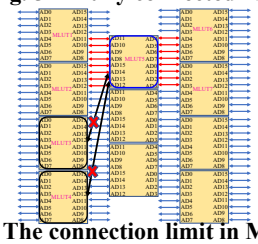


Fig. 6 The connection limit in MRLD

layer are fully connected with the preceding layer. For the MRLD structure, as shown in Fig. 6, each MLUT (such as MLUT5) can only connect up to four adjacent MLUTs (MLUT1, MLUT2, MLUT6, MLUT7), the data-outputs of the others MLUTs (such as MLUT3, MLUT4) cannot be connected to the MLUT (MLUT5). Thus, constructing a NN with the fully connection into the MRLD is impossible.

To implement a NN into MRLD, in this paper, we propose a sparse neural network based on the MRLD structure named MNN (MRLD-based Neural Network) as shown in Fig. 7 and Fig. 8, respectively. The proposed MNN has an inward gradual convergence and association characteristics in order to adapt the connection structure of MLUTs. 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. Fig. 9 shows an example of using MNN for a very simple image recognition application. Where, a 4×4 bit image of o and z is given respectively, and the vector of each row (4bit) is applied to the address input of an MULT at the first column of the MLUT array, respectively. Throughout the hidden layers, the feature of each row vector will be converged inwardly and gradually, and associated until the output layer, where all features will be extracted and recognized.

3.2 Implementing MNN into MRLD

Fig. 10 shows the wiring method to connect the neurons between adjacent layers in an MLUT array where each MLUT has 16 bits AD lines. The output of each neuron function configured in an MLUT will be read out and propagated to the following

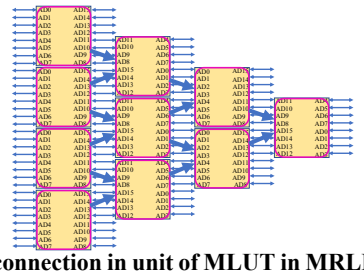


Fig. 7 Sparse connection in unit of MLUT in MRLD

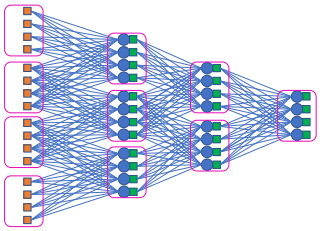


Fig. 8 Proposed MNN (MRLD-based Neural Network)

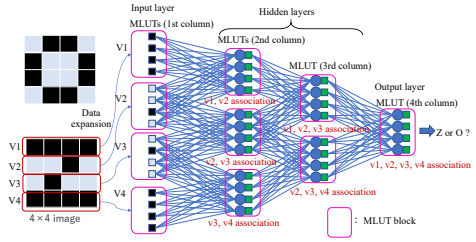


Fig. 9 Feature extraction in MNN

adjacent MLUT through only one AD (address-data) line. Then, the value of the address input from the preceding MLUT will be connected to the inputs of all neurons by configuring the branch logic (or wiring logic), e.g.: AD11 in MLUT x0y0 of Fig.10. Each neuron configured in an MLUT can connect to at most 8 neurons which are configured in the preceding adjacent MLUTs.

4. Experimental results

In this section, we describe the performed the experiments. First, we design an experiment to confirm the LUT-based neuron model as described in section 2. Then, we also show the experimental results to confirm the effectiveness of the proposed MNN in the section 3 by the training using the MNIST dataset.

4.1 Confirm the LUT-based neuron model

As shown in Fig .11 (a), here a size of 4×4×4 NN is given, and each layer (Hidden-layer1, Hidden-layer2, Output-layer1) is constructed to MLUT x0y1, MLUT x1y0, MLUT x2y1, respectively. For simplicity in this experiment, for this NN we using the *Heaviside Step* (*Binary step*) as an activation function:

$$f(u) = \begin{cases} 1 & ,u \geq 1 \\ 0 & ,u < 1 \end{cases}$$

$$u = \sum_{i=0}^{N} w_i \times x_i + b$$

to calculate each neuron. As shown in Fig. 11 (b), we give the value of the weights for each layer and assign the value of b to 0 . Each layer of given weights NN is calculated to a truth table stored in an MLUT to realize the neuron functions. Where, such as there are inputs 01010000, through the MLUT x0y1, MLUT x1y0, MLUT x2y1, theoretically the outputs of Hidden-layer1, Hidden-layer2, Output-layer is 0011, 1011, 0111, respectively. We also performed an logic simulation experiment in an MRLD with 16-bits 6×6 MLUTs array. As shown as Fig. 11 (c), the experimental result shows that the operating results of the LUT-

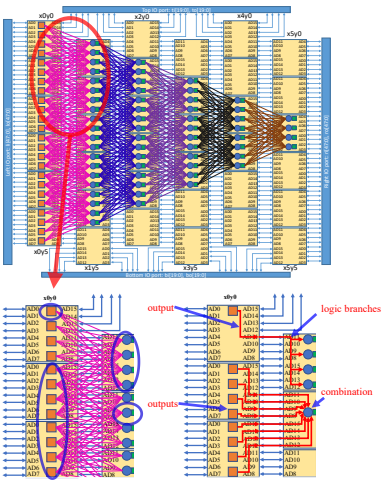
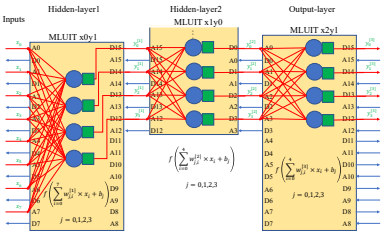
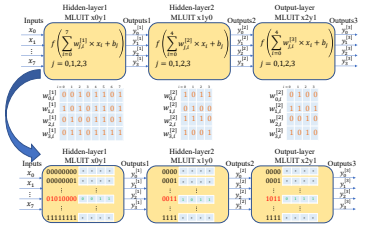


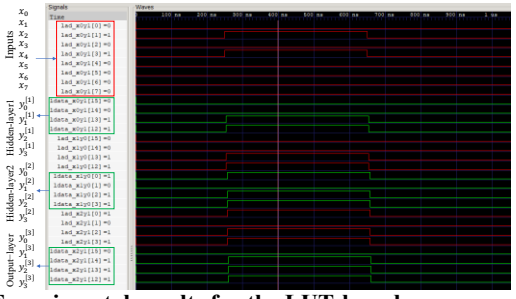
Fig. 10 The MNN wiring connection way in MRLD



(a) A size of 4×4×4 NN constructed in 3 MLUTs

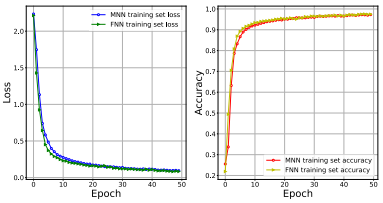


(b) LUT-based neuron model for the size of 4×4×4 NN

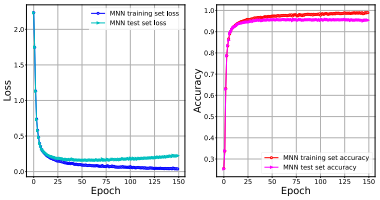


(c) Experimental results for the LUT-based neuron model

Fig. 11 Confirm the LUT-based neuron model



(a) The MNN and FNN training result in 50 epochs



(b) The MNN training result in 150 epochs

Fig. 12 The MNN confirmation results

based neuron model in MRLD are the same as the results of the above theoretical analysis.

4.2 Confirm the proposed MNN

In this experiment, for comparing with FNN, 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. 12 (a) shows the training results in 50 epochs, the results show that the MNN is an effective neural network that can get well accuracy and loss as same as the FNN. Fig. 12 (b) shows the MNN has been 150 epochs trained, and it can obtain the training accuracy and testing accuracy up to 0.99 and 0.96, respectively.

5. Conclusions

In this paper, we suggest a LUT-based neuron model to implement a NNs into MRLD device. The NN neuron's operation can be calculated into truth table form pre-stored in MLUT of MRLD. In MRLD, due to the special interconnection structure of MLUTs, it is difficult to construct a NN with fully connection into the MRLD. Therefore, we proposed a novel network structure MNN (MRLD-based Neural Network) to adapt the MRLD structure. To confirm the LUT-based neuron model, we design a logic simulation experiment by implementing a 4×4×4 LUT-based neural network. We confirm that the simulation results are the same as the results of the theoretical analysis. To evaluate the effectiveness of the MNN, we also 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," in *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," in *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] A. Wheeldon, R. Shafik, T. Rahman, J. Lei, A. Yakovlev, and O.-C. Granmo, "Learning automata based energy-efficient AI hardware design for IoT applications," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 378, no. 2182, p. 20190593, Oct. 2020.

[5] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, no. 1, pp. 239-255, 2010.

[6] Z. Fan, F. Qiu, A. Kaufman, S. Yoakum-Stove, "GPU Cluster for High Performance Computing," in *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*, O'Reilly, 2017.

[8] N. Singh, S. P. Panda, "Enhancing the Proficiency of Artificial Neural Network on Prediction with GPU," in *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," in *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," in *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] F. Piazza, A. Uncini and M. Zenobi, "Neural networks with digital LUT activation functions," *Proc. Int. Jt. Conf. Neural Networks (IJCNN)*, vol. 2, pp. 1401-1404, 1993.

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

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