Extended Abstracts of the 2015 International Conference on Solid State Devices and Materials, Sapporo, 2015, pp146-147

Memory-based LVQ Neural Network with Dedicated Learning Circuit

Xiangyu Zhang, Fengwei An, Lei Chen, and Hans Jürgen Mattausch

Hiroshima University, 1-3-1 Higashi-Hiroshima, Hiroshima 739-8530, Japan
Phone: +81-82-424-5730 E-mail: zhangxiangyu, anfengwei, chen, hjm@hiroshima-u.ac.jp

Abstract

This paper reports a dual-mode system, namely, on-chip learning based on learning vector quantization (LVQ), and recognition based on nearest-neighbor matching (NNM). The LVQ neural network algorithm is implemented on one chip using a pipeline with parallel p-word input (PPPI) architecture and a dedicated learning circuit. NNM, which results in high computational demand in both learning mode (LM) and recognition mode (RM), is solved by PPPI and reused for both modes so that a significant reduction in power consumption and area is achieved. The dual-mode system is highly flexible for used feature dimensionality and reference-vector number so that many different applications are implementable on this hardware platform. The fabricated test chip in 180 nm CMOS has parallel 8-word inputs, 102 K-bit on-chip memory, and achieves low power consumption of 66.38 mW at 75 MHz and 1.8 V supply voltage.

1. Introduction

Artificial neural networks (ANNs), which implement a simplified model of the human brain, specialize on pattern recognition, representing high-level processing in machine vision systems. Hardware implementations are necessary for ANNs in many practical applications to replace microprocessors, because they cannot handle the high computational costs. LVQ is a type of neural-network algorithm for which hardware implementations have already been realized as system-on-a-chip platforms [1], digital circuits [2, 3] and analog circuits [4]. Although these previous implementations provide massive intrinsic parallelism, adaptability to different applications is a still unsolved issue.

The proposed memory-based solution for LVQ with a dedicated learning circuit has large flexibility to cover many applications and also very low power consumption.

2. PPPI for LVQ1-based Learning and Recognition

LVQ1 refers to a group of algorithms applicable to statistical pattern recognition. Suppose that $x(t)$ and $w_s(t)$ represent an input sample and a winner reference vector in the discrete-time domain, respectively. $a(t)$ is the learning rate. Then $w_s(t)$ is updated to better comply with $x(t)$ according to following algorithm. Under the condition that $x(t)$ and $w_s(t)$ belong to same class, $w_s(t)$ is moved towards $x(t)$,

$$w_s(t+1)= w_s(t)+a[x(t)- w_s(t)] \quad (1)$$

In the other case, if $x(t)$ and $w_s(t)$ belong to different classes, $w_s(t)$ is moved away from $x(t)$,

$$w_s(t+1)= w_s(t)-a[x(t)- w_s(t)] \quad (2)$$

NNM is a common computational problem for many recognition and learning algorithms so that many advanced NNM techniques such as FPGA implementation of NNM [5], were proposed. To reduce the complexity of the NNM circuit, selection of the squared Euclidean distance (D_E^2) as distance metric is preferred, since the root operation has no influence on the accuracy. In case of d-dimensional input- and reference vectors, D_E^2 is described by (3).

$$D_E^{\ 2} = \sum_{i=1}^{d}(x_i - w_i)^2 \quad (3)$$

As shown in Fig.1, the PPPI architecture is composed of four parts: input layer, competition layer, winner-takes-all part, and output layer. The function variation of each part, which is enabled by mode control signal "L/R", multiplexers "M2", and de-multiplexers "DM", contributes to realizing the dual-mode. The input layer ensures the PPPI flexibility via partial vector storage which is described in more detail in [1]. The d-dimensional input vectors are placed into p memory blocks in the form of m partial vectors. Signal "Next", which distinguishes the distance calculation for two vectors, is asserted when the partial storage of one vector is finished. The competition layer, which executes the computing, consists of a weight unit and a summation unit, which are configured as a dedicated learning circuit in LM, and a distance accumulation adder tree in RM, respectively. The reference vectors are also partially stored in reference memories (REFs). In LM, the weight unit computes $[x(t)-w_s(t)]a$ and delivers the results to the dedicated learning circuit, which computes p updated components of the reference vector $w_s(t)+a[x(t)-w_s(t)]$ in parallel. Finally, the updated reference vector components are used to overwrite the old data in the REFs through the updating data bus. The sign of α is determined after class-label comparison between $x(t)$ and $w_s(t)$ through signal "C/I". In RM, the weight unit computes $[x-w_i]^2$ and delivers the results to the distance accumulation adder tree, which sums up all outputs from the weight unit so that D_E^2 is attained. The winner-takes-all part compares the intermediate minimum distance in register S5 to the accumulated D_E^2 in register S4, and then asserts a signal "Load" to the output layer if the distance in S4 is less than the intermediate minimum distance in S5. The output layer outputs the class label of $w_s(t)$ as the recognition result for $x(t)$.

In this research, to achieve pipelined learning, the REFs are implemented as 2-port memories with independent read-only and write-only ports. No special mechanisms are needed to detect and manage read-write conflicts in these 2-port memories because, in the PPPI architecture, the two

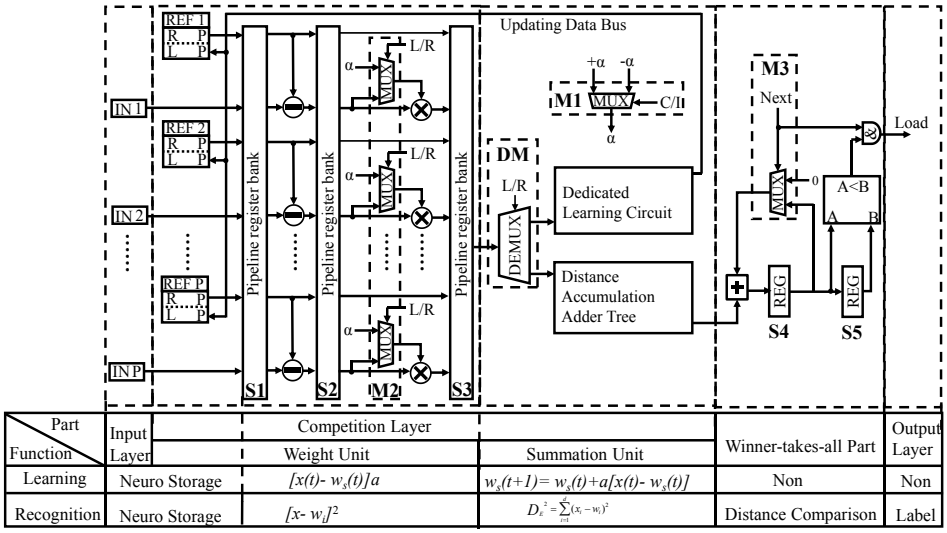

Fig. 1 PPPI architecture for a memory-based LVQ neural network. The same hardware parts are configured to have different functionality in different operating modes.

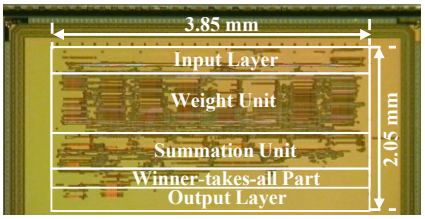

Fig.2 Micrograph of the fabricated chip in 180 nm CMOS technology with 8-word parallelism for the PPPI architecture.

Table I Performance Comparison

	[1]	This Work
CMOS technology	0.18µm	0.18µm
Power consumption (mW)	214 (25 MHZ @1.8V)	66.38 (75MHZ @1.8V)
Storage capability (K-bit)	96	102
S_r (µs)	0.28	0.106
S_l (µs)	20.9	1.15
Throughput (Gbits/s)	2.23	9.6

ports never access the same address at the same time which is guaranteed by a fixed delay between read and write operations.

3. Implementation Results and Conclusions

Fig.2 shows the fabricated test chip, which has a pipeline depth of 8 stages, an area of 7.89 mm^2, and 1024 max dimension of feature vectors. In the RM, the chip can process an 8-dimensional sample vector with a latency of 8 clock cycles (106 ns at 75 MHz) and a throughput of one 8 dimensional reference vector per clock cycle. The fabricated chip has highly flexibility of handling different reference vector dimensions so that it can handle a large number of different applications. For example, in the case of 120 dimensional reference vectors, there are m=15 partial vectors so that the recognition throughput becomes one reference vector in 15 clock cycles (200 ns at 75 MHz).

Comparison with previous state-of-the-art work is shown in Table I. The power consumption of the fabricated test chip is much smaller than in [1] even through our chip has more shared memory and higher operating frequency. As for efficiency, in the case of 8 or less than 8 dimensional reference vectors, both the minimal recognition speed (S_r) and the minimal learning speed (S_l) of each iteration for each reference vector are clearly better than in [1]. In addition, our test chip for conceptual verification of the developed memory-based LVQ architecture does not need any complex external control unit or a host PC.

In conclusion, the PPPI architecture for a memory-based LVQ neural network algorithm is verified to feature on-chip learning and recognition capability, to have very low power consumption, and also large flexibility for different applications.

Acknowledgements

This research was supported by grant 25420332 from the ministry of Science and Education, Japan. The VLSI-chip was fabricated through the chip fabrication program of VDEC, the University of Tokyo in collaboration with, Rohm, Synopsys, and Cadence. The used standard cell library was developed by Tamaru/Onodera Laboratory of Kyoto University and released by Professor Kobayashi of Kyoto Institute of Technology.

References

[1] F. An, T. Akazawa, S. Yamasaki, L. Chen, and H. J. Mattausch, *Japanese Journal of Applied Physics* **54.4s** (2015) 04DE05.

[2] M. Porrmann, U. Witkowski, and U. Rückert, *Neural Networks, IEEE Transactions on* **14.5** (2003) 1110.

[3] P. Ienne, P. Thiran, and N. Vassilas, *Neural Networks, IEEE Transactions on* **8.2** (1997) 315.

[4] L. M. Reyneri, *Neural Networks, IEEE Transactions on* **14.1** (2003) 176.

[5] F. An, and H. J. Mattausch, *Journal of Systems Architecture* **59.3** (2013) 155.