A VLSI Convolutional Neural Network for Image Recognition Using Merged/Mixed Analog-Digital Architecture

Keisuke Korekado[a], Takashi Morie[a], Osamu Nomura[b], Hiroshi Ando[c], Teppei Nakano[a], Masakazu Matsugu[b] and Atsushi Iwata[c]
[a] *Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, 808-0196, Japan*
E-mail: {korekado-keisuke@edu., morie@, nakano-teppei@edu.}brain.kyutech.ac.jp
[b] *Canon Research Center, Atsugi, 243-0193, Japan*
E-mail: {nomura.osamu, matsugu.masakazu}@canon.co.jp
[c] *Graduate School of Advanced Sciences of Matter, Hiroshima University, Higashi-Hiroshima, 739-8526, Japan*
E-mail: {ando, iwa}@dsl.hiroshima-u.ac.jp

Abstract: Hierarchical convolutional neural networks are a well-known robust image-recognition model. In order to apply this model to robot vision or various intelligent vision systems, its VLSI implementation with high performance and low power consumption is required. This paper proposes a VLSI convolutional network architecture using a hybrid approach composed of pulse-width modulation (PWM) and digital circuits. We call this approach merged/mixed analog-digital architecture. The VLSI chip includes PWM neuron circuits, PWM/digital converters, digital adder-subtracters, and digital memory. We have designed and fabricated a VLSI chip by using a 0.35 μm CMOS process. The VLSI chip can perform 6-bit precision convolution calculations for an image of 100×100 pixels with a receptive field area of up to 20×20 pixels within 5 ms, which means a performance of 2 GOPS. Power consumption of PWM neuron circuits is measured to be 20 mW. We have verified successful operations using a fabricated VLSI chip.

1. Introduction

For object detection or recognition from natural images, processing models for extracting image features should tolerate pattern deformations and pattern position shifts. Convolutional neural networks (CoNNs) with a hierarchical structure, which imitate the vision nerve system in the brain, have such functions [4, 7, 17, 8, 9].

The operations required for implementing convolutional networks are multiplication by weights and nonlinear conversion, as conventional neural network models. Because they require much computation, efficient VLSI implementation is needed for executing these operations in real-time and with low power consumption, in order to develop challenging intelligent applications such as robot vision. Various VLSI neural networks have actively been developed so far [13, 10, 2, 3, 18], and an analog VLSI processor suitable for convolutional networks was also reported [1]

On the other hand, we have already proposed a new circuit architecture, which is based on a pulse-width modulation (PWM) approach merging analog and digital approaches [6, 16, 15, 14, 11, 12, 5]. This architecture has various advantages of both approaches, especially it achieves low power consumption, and it is suitable for implementing neural networks.

This paper is organized as follows: In section 2, we describe a CoNN model for object detection or recognition. In section 3, we propose a VLSI CoNN architecture that consists of PWM neuron circuits and digital memory by combining the merged analog-digital architecture with the digital approach. In section 4, we propose a PWM neuron circuit that consumes very low power. In section 5, we present the measurement results of a VLSI chip fabricated using a 0.35 μm CMOS process. Finally, section 6 presents our conclusions.

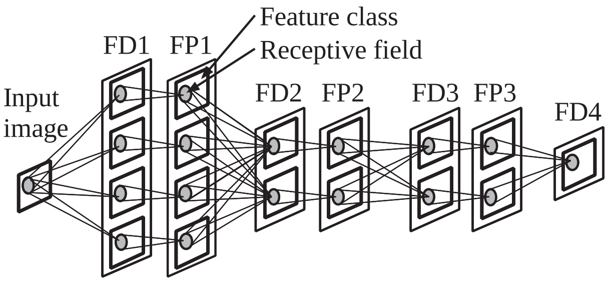

Figure 1: CoNN model for object detection or recognition

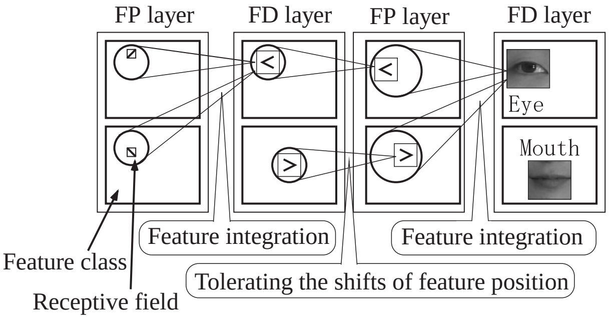

Figure 2: Principle of pattern detection using a CoNN (an example of eye pattern detection)

2. Hierarchical Convolutional Neural Network Model

The CoNN for object detection or recognition is shown in Fig. 1 [9]. The first layer of the hierarchical structure only receives images. Every following layer consists of two sub-layers: a feature detection (FD) layer and a feature pooling (FP) layer. Each layer includes some feature classes, each of which has neurons that react the same image feature. The neurons are arranged in a 2-D array to maintain the feature position of the input image. Therefore, the feature class pixel size is equal to the input image pixel size, and each neuron corresponds to each pixel. Each neuron is connected to the neurons located in a predefined area near the same position of the previous layer, which is called a receptive field.

Figure 2 shows the principle of pattern detection using a CoNN. The FP neurons are used to achieve recognition tolerant to pattern deformation and position shifts. The FD neurons operate for integrating a feature. By the hierarchically repetitive structure, local simple features (e.g., line segments) of the input image are gradually assembled into complex features.

Operations between layers are considered as a convolution because all neurons belonging to a feature class have a receptive field with the same weight distribution. The receptive field of the FP neuron is on the same feature class of the previous FD layer. All neurons of the FP layer have the same positive weight distribution, in which the weight is largest in the center of the receptive field and it decreases as the position is apart from the center. The shifts of feature positions in the FD layers are tolerated in the FP layers by this weight distribution. On the other hand, the receptive fields of the FD neurons are on all feature classes of the previous FP layer. The weights of the FD neurons are obtained by training.

Figure 3 shows the numerical simulation results of face position detection by a CoNN [9]. In this figure, brighter points indicate the positions of each feature. The FD1 layer detects line segments with various orientations of the input image. The FD2 layer combines the results of FD1 layer and detects intermediate features of moderate complexity, which are shaped like ‘$<$’ or ‘$>$’. In the similar way, the FD3 layer detects eyes and a mouth positions. By combining the results of FD3 layer, the FD4 layer detects the center position of the face.

3. VLSI Convolutional Neural Network Architecture

We propose a VLSI chip architecture that implements the CoNNs. The CoNN operation needs a huge number of multiplications for synaptic connections and the summations at each neuron. Because the number of processing circuits integrated in a chip is limited, it is difficult to construct all connections of the hierarchical network by real processing circuits. Therefore, in our architecture, neuron circuits are repetitively used by time-sharing operation.

Time-sharing operation in the CoNN is shown in Fig. 4. The feature class size and the receptive field size are assumed $N \times N$ and $m \times m$ pixels, respectively. The previous layer has $N_P \times N_P$ $(= (N+m-1) \times (N+m-1) > N \times N)$ neurons, in which the extra neurons of the exterior of the feature class area output a zero value, so that the feature class size is the same in all the lay-

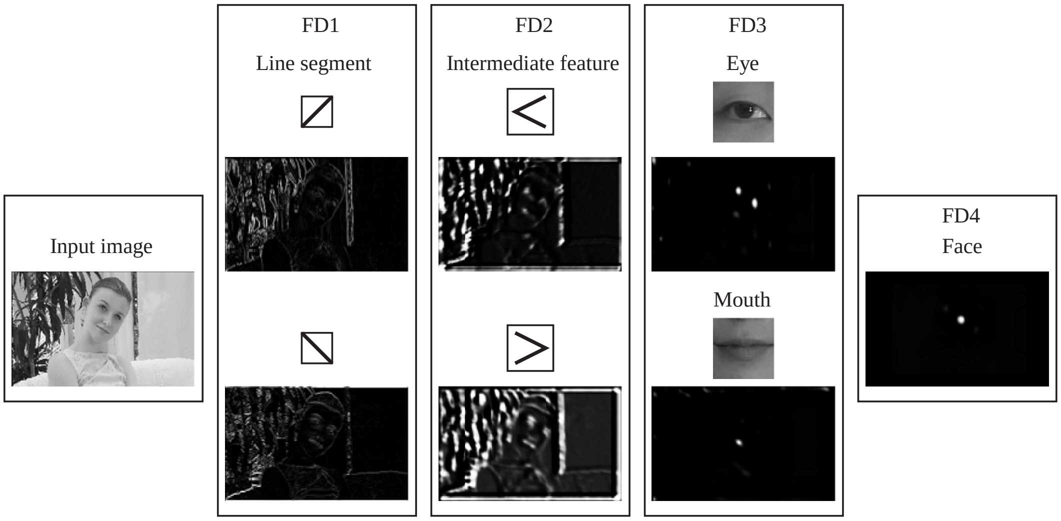

Figure 3: Numerical simulation results of face position detection by a CoNN

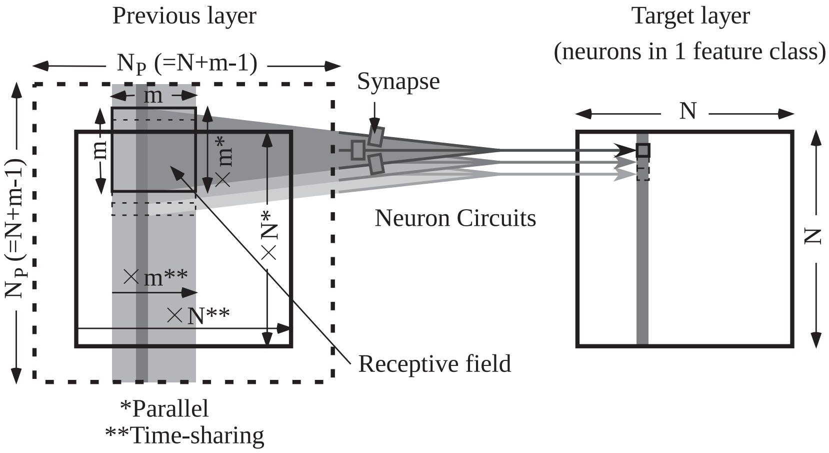

Figure 4: Time-sharing operation in the CoNN

ers. By these extra neurons of the previous layer, we can detects the center position of a face even if the input image does not have the complete face image. The outputs of N_P neurons belonging to one column of the previous layer are inputted to N neuron circuits simultaneously. The neuron operations for one column of the receptive fields are performed in parallel. For m rows in a receptive fields, the above neuron operations are repeated by m times, and furthermore each of them is repeated twice for positive and negative weighting. Thus, the number of repetitions in N-parallel neuron operations for convolution between feature classes is $N \times m \times 2$.

The block diagram of our CoNN circuit is shown Fig. 5. By utilizing the advantage of small circuit size in the PWM approach, m-input PWM neuron circuits are integrated. To achieve time-sharing operation, the partial accumulation results of neuron operation are temporarily held in the neuron circuit. These partial results are accumulated and stored in an SRAM through the PWM/digital converter (WDC) and the digital adder-subtracter (DAS). The WDC converts PWM signals output from a neuron circuit into digital signals. The DAS is used in time-sharing operations for one column of the receptive field and for the positive and negative weighting.

Although we assumed that the number of inputs of the neuron circuits is $m \times m$, convolution with a smaller receptive field size can be calculated by setting the extra inputs at zero. Convolution with a larger receptive field size can also be calculated by time-sharing operation.

4. PWM Neuron Circuit

4.1. *Connection Model*

In the general feedforward networks, internal state u_i and output o_i of postsynaptic neuron i are given by the following equations, respectively;

$$u_i = \sum_j w_{ij} o_j \ , \tag{1}$$

$$o_i = f(u_i) \ , \tag{2}$$

where w_{ij} is the connection weight from presynaptic neuron j to postsynaptic neuron i, and f is the nonlinear conversion function.

In the conventional model, the synapse part multiplies o_j by w_{ij} and the neuron (soma) part executes summation and nonlinear conversion $f(u_i)$, as shown in Fig. 6(a). From eqs. (1) and (2), the output of postsynaptic neuron i is given by

$$o_i = f(\sum_j w_{ij} o_j) \ . \tag{3}$$

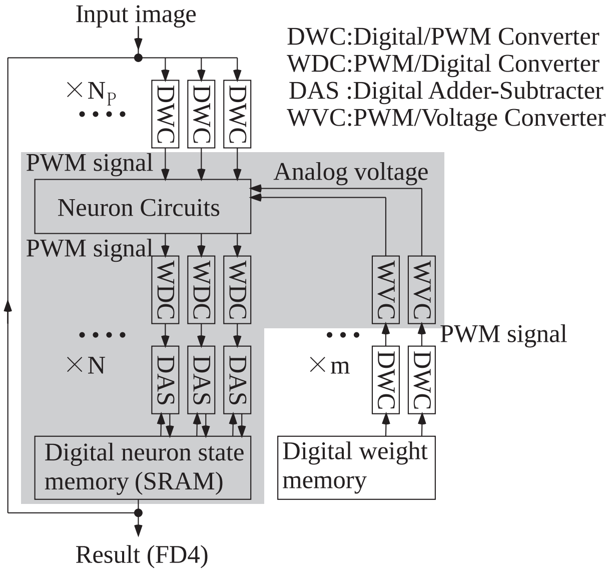

Figure 5: Block diagram of our CoNN circuit. The components in the shaded region are included in our VLSI chip

However, in our circuit model, both multiplication and nonlinear conversion are performed by the synapse part, and the neuron part executes summation and outputs the internal state, as shown in Fig. 6(b). Thus, from eqs. (1) and (2), the internal state which is the output of neuron i is given by

$$u_i = \sum_j w_{ij} f(u_j) \ . \tag{4}$$

Equations (3) and (4) are the equivalent operations in hierarchical networks.

4.2. *Circuit Design*

Our PWM neuron circuit includes plural synapse circuits, which operate in parallel, and an integrating capacitor as shown in Fig. 7. Its operation is as follows: (1) A PWM signal U_j that corresponds to the internal state of the presynaptic neuron is transmitted to the

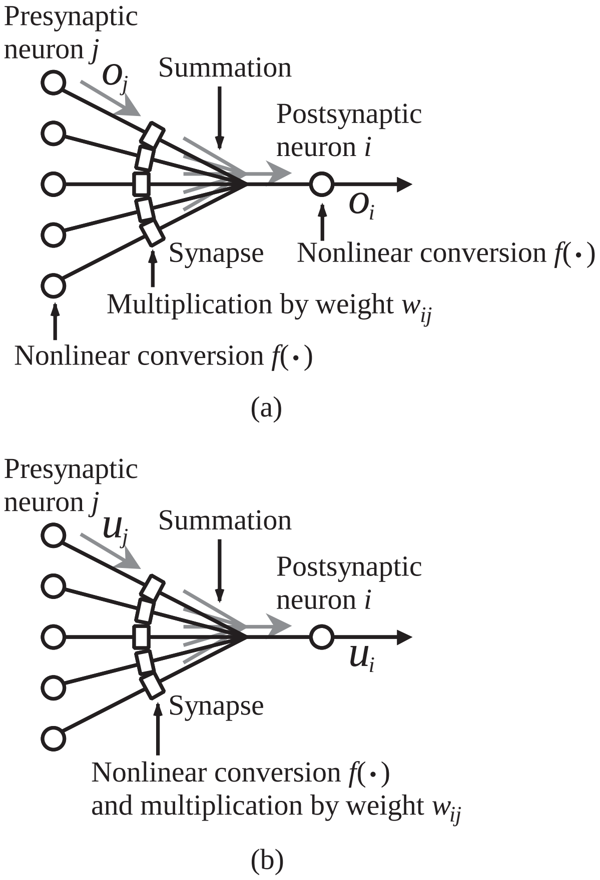
Presynaptic
neuron j
o_j
Summation
Postsynaptic
neuron i
o_i
Synapse
Nonlinear conversion $f(\cdot)$
Multiplication by weight w_{ij}
Nonlinear conversion $f(\cdot)$
(a)
Presynaptic
neuron j
u_j
Summation
Postsynaptic
neuron i
u_i
Synapse
Nonlinear conversion $f(\cdot)$
and multiplication by weight w_{ij}
(b)

Figure 6: Connection model: (a) ordinary model, and (b) our neuron circuit model

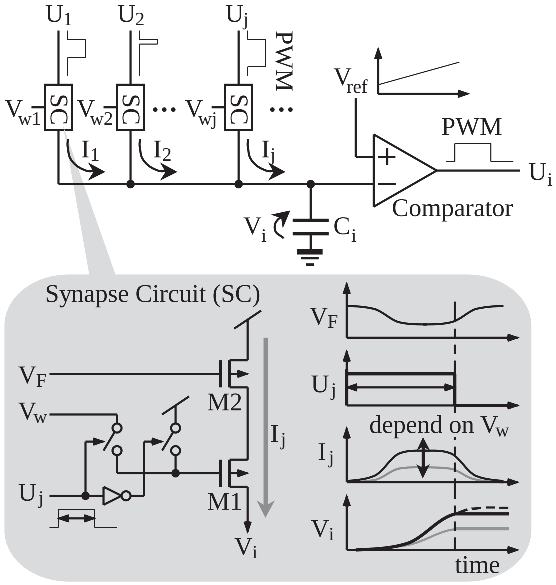

Figure 7: PWM neuron circuit

synapse part; (2) the input PWM signals are converted with the nonlinear function, and weighted summation is performed by converting the PWM signals into charges stored in capacitor C_i; (3) the voltage between the nodes of the capacitor, V_i, is converted into a PWM signal by comparing it with linearly-ramped voltage signal V_{ref}.

In this circuit, connection weighting (multiplication) and nonlinear conversion are performed by two MOSFETs, *M1* and *M2*, at the same time. The connection weighting is achieved by applying analog DC voltage V_w, which controls the conductance of *M1*. The nonlinear function is applied to all synapses by changing analog voltage V_F, which controls the conductance of *M2*.

The waveform of V_F as a function of time has the similar shape to the differential function of the nonlinear conversion function f although the waveform of V_F is reversed because *M2* is a p-type MOSFET. The additional charges stored in capacitor C_i is determined by integrating current I_j flowing to C_i during the time span defined by the pulse width of the input PWM signal. As V_F lowers slightly beyond the threshold voltage of *M2*, current I_j smoothly increases from zero depending on V_F, because *M2* operates in the saturation region. Thus, capacitor voltage V_i changes smoothly. (The same situation also occurs when V_F rises near the threshold voltage of *M2*). When V_F lowers further, *M2* operates in the triode region although *M1* still operates in the saturation region. Therefore, in the time span when I_j is large, current I_j depends on V_w. Thus, weighting and nonlinear conversion $w_{ij} \cdot f(u_j)$ are achieved at the synapse part.

The PWM neuron circuit consumes much lower power than the corresponding digital circuit, which we can construct from digital multipliers and look-up tables, because the PWM neuron circuit performs multiplication and nonlinear conversion by only one switching operation.

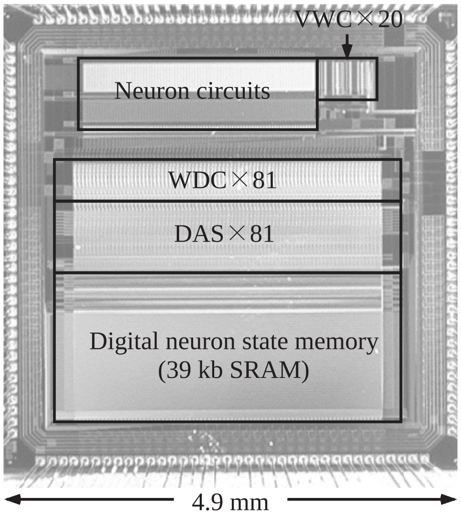

Figure 8: Chip micro-photograph

5. Experimental Results Using a Fabricated VLSI Chip

We fabricated a CoNN VLSI chip by using a 0.35 μm CMOS process based on our architecture. Figure 8 shows a micro-photograph of the fabricated chip. The VLSI chip includes 81 neuron circuits with 20 synaptic inputs, 81 PWM/digital converters, 81 digital adder-subtracters, 39 kb SRAM, and 20 weight setting circuits. Therefore, this chip can implement a CoNN with $N_P = 100$ and $m = 20$. By using this VLSI chip with external feedback control, we can construct hierarchical structure of a CoNN.

We measured the PWM input-output relationship of neuron circuits when all of 20 PWM input signals per neuron are identical. The measurement results are shown in Fig. 9 with the corresponding circuit simulation (HSPICE) results. In this figure, V_w is changed from $1.6\ V$ (the largest weight) to $2.7\ V$ (the smallest weight) at $0.1\ V$ intervals. Although the measurement values are slightly larger than the simulated ones, the characteristic tendency agrees well with the simulation results, and the differences can be adjusted by changing V_w. Therefore, it is demonstrated that weighting and nonlinear conversion are achieved simultaneously.

We measured the convolution operation when we input the neuron states and weight shown in Fig. 10(a). The measurement result is shown in Fig. 10(b) with the theoretical one. We obtained the vertical pattern edges of the neuron states by using the weight distribution of the receptive field. The measurement result agrees well with the theoretical one, and it is demonstrated that correct convolution operation is achieved. We also verified that all circuit components operate successfully.

We defined the operation cycle time as 1.6 μs. Under this condition, the chip achieves an operation performance of 2 GOPS[1] by parallel operations for 81 $(= N)$ neurons and 1620 $(= N \times m)$ synapses, since this chip achieves 1620 $(= N \times m) \times 2$ (multiplication and summation) operations per operation cycle 1.6 μs. The performance of 2 GOPS can also be expressed as 1 GCPS[2]. The whole convolution operation requires about 5 ms, since the number of repetitions for convolution between feature classes is 3240 $(N \times m \times 2)$ as we described in section 3.

We measured a power consumption of PWM neuron circuits to be 20 mW although the digital circuit block consumes 100 mW (WDC and DAS consume power of 70 mW, SRAM consumes power of 30 mW). In order to reduce the total power consumption of the chip, lower-power digital circuit design, especially redesign of the WDC is an important future issue.

6. Conclusions

We proposed a merged/mixed analog-digital VLSI chip architecture for CoNNs using PWM and digital circuit techniques.

A neuron circuit with 20 synapses was designed. Nonlinear conversion and multiplications by connection weights are realized by two MOSFETs, thus a very small layout area and low power consumption of the synapse part were achieved. Since the connections between layers have the same weight distribution, hierarchical networks can be constructed by feedback and time-sharing operations using the VLSI CoNN.

We designed and fabricated a VLSI CoNN with an operation performance of 2 GOPS, and verified successful operations of all circuit components.

Acknowledgements

The VLSI chip has been fabricated in the chip fabrication program of VLSI Design and Education Center

[1] Giga Operations Per Second.
[2] Giga Connections Per Second.

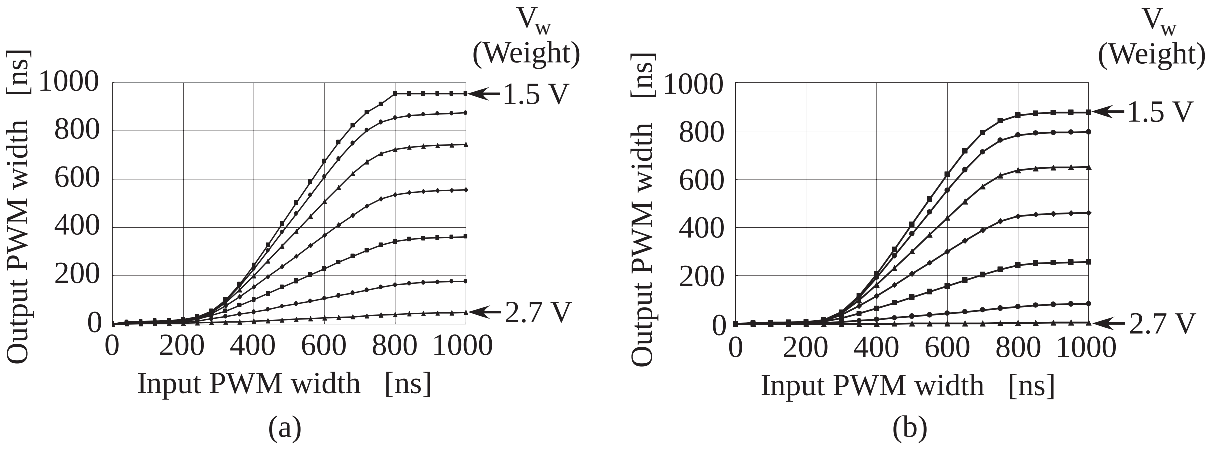

Figure 9: PWM input-output relationship: (a) measurement results, and (b) circuit simulation (HSPICE) results

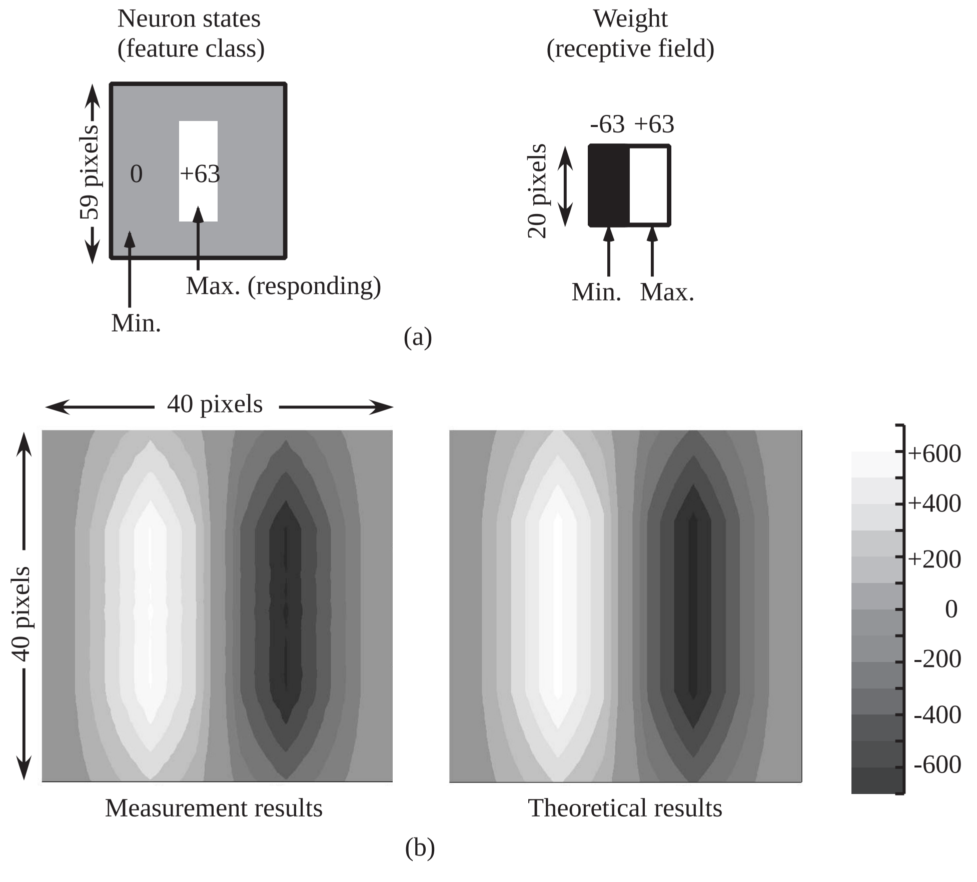

Figure 10: Convolution operation: (a) input neuron states and weight, and (b) measurement results of convolution operation and theoretical results

(VDEC), the University of Tokyo, Japan with the collaboration by Rohm Corporation and Toppan Printing Corporation.

References

[1] B. E. Boser, E. Säckinger, J. Bromley, Y. Le Cun, and L. D. Jackel, *An analog neural network processor with programmable topology*, IEEE J. Solid-State Circuits **26** (1991), no. 12, 2017–2025.

[2] G. Cauwenberghs and M. A. Bayoumi (eds.), *Learning on silicon: Adaptive VLSI neural systems*, Kluwer Academic, Norwell, MA, 1999.

[3] L. Chen and B. Shi, *CMOS PWM VLSI implementation of neural network*, Proc. Int. Joint Conf. on Neural Networks (IJCNN), vol. III, 2000, pp. 485–488.

[4] K. Fukushima and S. Miyake, *Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position*, Pattern Recognition **15** (1982), 455–469.

[5] A. Iwata, T. Morie, and M. Nagata, *Merged analog-digital circuits using pulse modulation for intelligent SoC applications*, IEICE Trans. Fundamentals. **E84-A** (2001), no. 2, 486–496.

[6] A. Iwata and M. Nagata, *A concept of analog-digital merged circuit architecture for future VLSI's*, IEICE Trans. Fundamentals. **E79-A** (1996), no. 2, 145–157.

[7] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, *Face recognition: A convolutional neural-network approach*, IEEE Trans. Neural Networks **8** (1997), 98–113.

[8] M. Matsugu, *Hierarchical pulse-coupled neural network model with temporal coding and emergent feature binding mechanism*, Proc. Int. Joint Conf. on Neural Networks (IJCNN), 2001, pp. 802–807.

[9] M. Matsugu, K. Mori, M. Ishii, and Y. Mitarai, *Convolutional spiking neural network model for robust face detection*, Proc. Int. Conf. on Neural Information Processing (ICONIP), 2002, pp. 660–664.

[10] T. Morie and Y. Amemiya, *An all-analog expandable neural network LSI with on-chip backpropagation learning*, IEEE J. Solid-State Circuits **29** (1994), no. 9, 1086–1093.

[11] T. Morie, J. Funakoshi, M. Nagata, and A. Iwata, *An analog-digital merged neural circuit using pulse width modulation technique*, IEICE Trans. Fundamentals. **E82-A** (1999), no. 2, 356–363.

[12] T. Morie, M. Miyake, S. Nishijima, M. Nagata, and A. Iwata, *A multi-functional cellular neural network circuit using pulse modulation signals for image recognition*, Proc. Int. Conf. on Neural Information Processing (ICONIP) (Taejon, Korea), Nov. 2000, pp. 613–617.

[13] A. F. Murray and P. J. Edwards, *Enhanced MLP performance and fault tolerance resulting from synaptic weight noise during training*, IEEE Trans. Neural Networks **5** (1994), no. 5, 792–802.

[14] M. Nagata, J. Funakoshi, and A. Iwata, *A PWM signal processing core circuit based on a switched current integration technique*, IEEE J. Solid-State Circuits **33** (1998), no. 1, 53–60.

[15] M. Nagata and A. Iwata, *PWM signal processing architecture for intelligent systems*, Computers & Electrical Engineering **23** (1997), no. 6, 393–405.

[16] M. Nagata, T. Yoneda, D. Nomasaki, M. Sano, and A. Iwata, *A minimum distance search circuit using dual-line PWM signal processing and charge packet counting techniques*, IEEE Int. Solid-State Circuits Conf. Dig., 1997, pp. 42–43.

[17] C. Neubauer, *Evaluation of convolutional neural networks for visual recognition*, IEEE Trans. Neural Networks **9** (1998), 685–696.

[18] J. Schemmel, F. Schürmann, S. Hohmann, and K. Meier, *An integrated mixed-mode neural network architecture for megasynapse ANNs*, Proc. Int. Joint Conf. on Neural Networks (IJCNN), 2002, pp. 2704–2709.