An Area Efficient Shared Synapse Cellular Neural Network for Low Power Image Processing

Jinwook Oh, Seungjin Lee, Joo-Young Kim, and Hoi-Jun Yoo
Division of Electrical Engineering, School of EECS, KAIST, Daejeon, Republic of Korea
jinwook@eeinfo.kaist.ac.kr

Abstract—This paper presents an area and power efficient cellular neural network (CNN) that enables real-time image processing. The proposed shared synapse architecture halves the number of required synapse multipliers, which are the main contributor to area and power consumption of CNNs. For this, a current holder circuit is used to sample and hold the currents of non-changing synaptic circuit outputs. Compared to the conventional architecture of CNNs, power and area are reduced by 46% and 41%, respectively.

I. INTRODUCTION

Cellular neural network (CNN) has been studied for real-time image processing due to their fast operation, scalability, and feasibility for VLSI implementation [1, 2, 3]. In CNNs, each cell performs its own computation in parallel and is fully connected to adjacent cells so that all required operands are accessed simultaneously. The other advantage of CNN is its feasibility in VLSI implementation. Since each CNN cell only accesses directly adjacent cells, the number of required on-chip wires doesn’t exceed the practical level and adding more cells doesn’t affect existing CNN cells.

A CNN consists of a 2-dimensional regular array of identical cells. In Fig.1, each cell has its own generic transfer functions realized as the control template, the feedback template and the threshold template. These templates require one multiplier for each of the 19 connections to adjacent cells. Typically the multiplier of a synapse is implemented with linearized analog multipliers, operational transconductance amplifiers (OTA), and current scaling circuits to minimize nonlinear errors and offsets [4, 5]. In order to avoid mismatch error and noise, large transistors and/or differential architectures are needed, leading to increased area and power consumption. As a result, large area and power are consumed on the template realizations of conventional CNNs. In order to reduce the area required by the synapses, circuits using a minimal number of transistors such as the one transistor synapse [6] or the ohmic-MOST synapse [7] have been proposed. However, the actual area reduction is only incremental due to the large size of each individual transistor.

This paper presents the shared synapse CNN, which reduces area and power consumption of CNN cells while retaining functionality and accuracy. This is accomplished by exploiting the dynamics of the CNN cell. As Fig. 1 illustrates, two templates of CNN that normally require eighteen synaptic circuits are unified using only nine synaptic circuits accompanied by configurable interconnection and the proposed current holder. The current holder replaces nine of the synaptic circuits by sensing and holding the sum of pre-evaluated template output currents. The proposed circuit achieves 46% power reduction and 41% decrease in area consumption compared to the conventional 19-multiplier CNN cell.

II. SHARED SYNAPSE CNN CELL ARCHITECTURE

The conventional CNN cell operation is described by the following equation:

\[
\frac{dx(t)}{dt} = -x(t) + \sum_{a \in A} a_i y(t) + \sum_{d \in D} b_d u^d + i
\]

\[
f(x) = \frac{1}{2} \{|x - 1| + |x + 1|\}
\]
The $u^d$, $y^d$, and $i$ denote the input, output and threshold of each cell, respectively. $a_j$, $b_j$ are the coefficients of the feedback template $A$ and control template $B$. $f(x)$ is a piecewise linear function that determines final output. Since the term $\sum_{d \in N_r} b_d u^d$ in (1) is not affected by the feedback of the adjacent cell values, this term can be considered as static value once it is initially calculated. This characteristic of the CNN lead us to propose the shared synapse cell that performs CNN operation with only half number of multipliers compared to the conventional 19 multiplier CNN cells. The operation of the proposed CNN can be explained with two phases:

1. Intermediate term = $\sum_{d \in N_r} b_d u^d$  

2. ($\phi_1$) ($\phi_2$) Interm. term

In the proposed CNN, operation of each cell is divided into two phases to use synapses in time-multiplexed fashion. In the first phase, initial multiplication of external inputs and control template coefficients are performed to compute static intermediate term (3). And then, feedback template coefficients are multiplied with cell state values to yield final output value (4).

Fig. 2 shows the proposed shared synapse CNN architecture and its operation. To store pre-calculated term $\sum_{d \in N_r} b_d u^d$ as a static current, a current holder is proposed and it replaces 9 multipliers that compute the term $\sum_{d \in N_r} b_d u^d$. To use the same multipliers for control and feedback template calculations, both linear and non-linear voltage-current (V/I) conversions should be performed in the same cell. Therefore, we adopted a gm-cell to control linearity. The configurable interconnections between the input, the synapse and the cell memory are realized with nMOS switches. Since the signal range of the cell state is distributed from 0.5V to 1.3V, nMOS switch can be used in this circuit without any voltage degradation.

As shown in Fig. 2, the operation of the proposed architecture is divided into two mutually exclusive phases, denoted in $\phi_1$ and $\phi_2$. The interconnections of the current holder as well as the synapses are dynamically configured according to the phase that is active. $\phi_1$ is the sensing phase in which the sign and magnitude of the aggregate current output from the control template in the shared synapse is stored by the current holder. This is accomplished by storing the gate voltage of the current sensing transistor in the sensing capacitor. $\phi_2$ is the current holding phase. During this phase, the sensing capacitors are disconnected from the sensing transistors and are connected to the output driving transistors to duplicate the current of the control template. Meanwhile, the weights and inputs of the synapse circuits are configured for feedback template operation. As a result, the output currents from the synapse circuits and current holder are accumulated by the cell memory to achieve the CNN operation described by (4).

III. CIRCUIT IMPLEMENTATION

A. Current Holder

Fig. 3 shows the proposed current holder schematic. When the input current $I_{\text{control}}$ is sensed, the gate-source voltage of sensing transistors $M_1$ or $M_2$ is stored to a sensing capacitor $C_1/C_2$, since only one of the two current paths in the input column can be activated depending on the direction of the input current. These capacitors enable the output control
column to provide the sensed $I_{\text{control}}$ to next circuit by $M_3/M_4$. Then the input column is connected to the output column of feedback. The hold current for control template is determined as the sum of $I_{\text{control}}$ and $I_{\text{offset}}$ which exists between VDD and GND.

There are two important design issues that can affect CNN operation severely. The existence of $I_{\text{offset}}$ disturbs a sensing transistor $M_1/M_2$ to store the precise gate voltage at the sensing capacitor. This current can be several hundreds nA that induces more than 10% inaccuracy in holding output current and also consumes undesirable extra power. This problem is resolved with stacked transistors as 4 cascode transistors in Fig. 3 that remove the direct current path between VDD and GND. Then the error current is decreased to less than 10nA which is negligibly small.

Another error source is the clock feedthrough invoked at the switching moment. Mainly, the turn-off process of the switch next to the capacitor $C_1/C_2$ incurs the voltage variation, therefore, the hold current is changed with respect to the voltage error. In Fig. 4, the simple solution of this problem is presented. By placing a MOS capacitor which has the half size of the switching transistor, the voltage variation through the parasitic capacitor $C_p$ can be reduced effectively. Additionally, it can also eliminate the error caused by the charge injection to the storage capacitor $C_s$. In this design, 0.1V clock feedthrough error is compensated to less than 1% of switched signal swing.

B. Dual-mode V/I converter

In Fig. 5 (a), V/I converters are employed with a basic gm-cell to control linearity. This component is also shared in both control and feedback template multiplications by changing its linearity adaptively. $V_{\text{comp}}$ determines the center of the V/I curve and $V_{\text{cont}}$ changes the linearity of this circuit with a MOS transistor. When this circuit is set for linear mode operation, the slope of this curve is decreased. Therefore, the large $g_{m1}$ of input transistor $M_1$ and $M_2$ is designed not to distort CNN operation under the signal range from 0.5V to 1.3V.

C. Synaptic multiplier

Due to linear mode operation and small area consumption, the current scaling synaptic circuit is used for template realization as Fig. 5 (b). Using transistors $M_1$ and $M_2$ in triode mode, the linearity is achieved with small area and power consumption compared to the OTAs. In order to attain more accuracy, the W/L of mirroring transistor $M_3$ and $M_4$ should be large enough to hold $V_{ds}$ of $M_1$ and $M_2$ robustly for voltage to current relation in linear region.
IV. IMPLEMENTATION RESULT

Fig. 6 shows the simulation waveforms of the evolution in the proposed CNN. For the template, the connected component detector (CCD) is selected for the 3 unit cells with 2 boundary cells. By multiplying feedback template and propagating the data to neighbors, the state values will converge to the settling points finally. For the proposed CNN, the settling time is around 160ns.

While the conventional CNN requires only synaptic multiplier operation time, the proposed CNN additionally requires the charging time in sensing and hold operation for the shared synapse. However, its 5ns charging time is negligibly small in the whole CNN operation. Including CCD, the CNN was successfully tested for several template setting such as edge detection, hole filling, and erosion.

Base on the proposed shared synapse architecture, 80x60 CNN cell array is implemented for the real-time image processing and the cell is shown in Fig. 7. Not only area but also power consumption is reduced because the current hungry synapses are shared with the help of the proposed current holder. As a result, a cell of the proposed CNN occupies 910μm² area and dissipates 53μW, which are reduced by 41% and 46% compared to the conventional Chua-Yang model [1], respectively.

Table I. Shared Synapse CNN vs. Other CNN Architectures

<table>
<thead>
<tr>
<th>Process (μm)</th>
<th>[5]</th>
<th>[5]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply voltage (V)</td>
<td>3.3</td>
<td>3.3</td>
<td>2.5</td>
</tr>
<tr>
<td>Cell size (μm²)</td>
<td>120x102</td>
<td>73x76</td>
<td>&lt;244x244</td>
</tr>
<tr>
<td>Cell density (Cell/mm²)</td>
<td>82</td>
<td>190</td>
<td>167</td>
</tr>
<tr>
<td>Power per cell (μW)</td>
<td>370</td>
<td>180</td>
<td>375</td>
</tr>
<tr>
<td>τ (μs)</td>
<td>1.2</td>
<td>0.8</td>
<td>4.6</td>
</tr>
<tr>
<td>No. of synapses</td>
<td>9, 9, 2</td>
<td>9, 1, 1, 1</td>
<td>5, 5, 1</td>
</tr>
<tr>
<td>Weight distribution</td>
<td>76 b</td>
<td>6 b</td>
<td>[4, 4]</td>
</tr>
</tbody>
</table>

In Table I, the reported CNN structure is compared with reference designs. Although considering other peripheral circuits, the proposed architecture still achieves high compactness and power-efficiency with 4 times smaller and 3 times efficient than conventional CNN circuits. In addition, the smallest time constant of 66ns is obtained from the reduced architecture and increased density.

V. CONCLUSION

We present a shared synapse CNN for small area and low power consumption real-time image processor in 0.18μm CMOS technology. The proposed current holder and configurable synapses are designated for the shared synapse CNN. As a result of merged template multiplications in one synapse array, this architecture reduces its power and area overhead up to 46% and 41%, respectively, compared to the conventional CNN model.

REFERENCES