In-sensor time-domain classifiers using pseudo sigmoid activation functions

Ethan Chen, Vanessa Chen *

Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 15213, USA

Keywords: Classifier, Machine learning, Smart sensors, Time-domain, Inner-products

ABSTRACT

This work presents an ultra-low-power classifier that can be integrated within energy-constrained bio-sensors to enable rapid analysis for continuous health monitoring. The in-sensor classifier saves significant transmission energy by extracting critical information locally to eliminate the need of transmitting raw data to centralized servers for remote signal processing. The convolutional-neural-network (CNN)-based classifier is built by using reconfigurable delay-locked loops (DLLs) to carry out classification algorithms with time-domain multiply-accumulate (MAC) operations. Pseudo sigmoid activation functions are realized by regenerative comparators that transform weighted timing to probabilities. The presented classifier achieves low-power consumption of 240.34 nW while performing up to 20 k operations per second. The proposed time-domain classifier reduces the energy to 36% of the previous works.

1. Introduction

To continuously monitor health conditions, distributed sensors are designed to capture and transmit psychological signals, such as electro-cardiogram (ECG) or electroencephalogram (EEG), to the cloud for anomaly analysis, which is of great clinical importance. For example, hypertension accounts for about 25% of heart failure cases [1]. Real-time monitoring can be utilized to predict the emergency cases and diagnose the diseases before they become worse. That brings new challenges for pervasive edge sensors to enable the always-on feature for real-time tracking because transmitting raw data of the acquired signals to the aggregator burns a tremendous amount of energy. Comparing to full-waveform transmission, in-sensor computing or machine learning can be performed at edge sensors to extract critical features in situ and that further reduces volumes of transmission data [2–4]. In this way, only classified results will be sent to the aggregator, so transmission energy can be highly decreased to enable continuous monitoring.

Sensory interfaces to acquire EEG or ECG signals usually require more than 16-bit resolution [5]. High-performance analog-to-digital converters (ADCs) are often used to convert captured signals to digital data for digital signal processing (DSP). Automatic bio-signal analysis with statistical learning has been utilized for several years. Those digital architectures can be used as powerful accelerators [6–8]. However, machine learning operations are computationally expensive with modern computing systems for edge sensors. Moreover, they all require data converters, including ADCs and digital-to-analog converters (DACs) to interface with the sensors [9,10]. Recently, computational transformation can be embedded into ADCs to execute multiplying operations.
and to complete classification with backend processing [11–13]. To emulate biological sensory systems that are considered the most energy-efficient computers with analog signal processing [14], this paper utilizes CNN to enable direct classification in analog domain without sending data to or retrieving data from central processing units (CPUs) through data conversion to enhance data movement.

In order to decrease the energy consumption to its limits, lowering power supply voltages is an efficient approach. In this way, bio-sensors may use the energy harvested from the environment with the lowest maintenance [15]. As the CMOS technology is scaled down, the power supply is also scaled down to prevent the gate oxide from breakdown. While technology scaling with improved power and performance characteristics has brought tremendous benefits to digital circuits, the analog circuit design is becoming challenging due to the reduced intrinsic gain and limited headroom. Representing signals in time domain to achieve required resolution is beneficial because the unit delay of minimum-sized devices becomes finer with scaling. Hence, processing signals in time domain overcomes the difficulties of signal processing in voltage domain.

The presented approach focuses on signal processing in time domain to address low-headroom issues so that time-domain classification can be performed under low supply voltages to achieve better energy efficiency and benefit from technology scaling. Nevertheless, the greatest benefit along with the technology scaling is the increase of transit frequency and the decrease of propagation delay. Excellent timing accuracy is easily achieved when the transition time reaches the order of less than 10 ps? Meanwhile, smaller parasitic capacitance which comes from the smaller transistor size can decrease transition energy. Therefore, in order to get the most benefits from the progress of processes, digitizing more mixed-signal blocks to operate in time domain is an efficient method [16, 17].

2. Time-domain classifier

Extracting all of the features is very power consuming and impossible to be realized in the edge sensors. To achieve lower power consumption, mixed-signal classification structures with primary feature extraction...
shown in Fig. 1 exploit time-domain multiplication and summation to perform the following inner-products

\[ V_{\text{out}} = \sum_{i=1}^{N} (W_i X_i) \]  

(1)

Pseudo-sigmoid activation functions that are generated with regenerative comparators [18] calculate the likelihood for forward propagation of signals. A multi-layer neural network for classification with the proposed pseudo-sigmoid function is adopted in the paper. As the number of neurons grows to more than 200, the mean squared error increases significantly. Therefore, the structure with 100 neurons at each layer for 2 hidden layers is utilized with considering the accuracy and hardware overhead. Offline training to derive the weights is employed for further reduction of power consumption. Although rectified linear activation unit (ReLU) is popular in the CNN implementation recently because it is simple and can be easily implemented in the software-oriented classification. However, to implement it in the sensors in analog domain, it needs amplifiers with the closed-loop configuration to accomplish the linear part. The closed-loop amplifiers need large power consumption and supply headroom to achieve high linearity. It would lead to difficulties to integrate the classifiers into sensor front-end circuits, especially in advanced technology nodes. Therefore, the pseudo-sigmoid activation function is utilized in the proposed structure for better integration. The problem of vanishing gradients that nonlinear activation functions encounter in deep neural networks does not cause problems in this structure since the adopted structure only contain 2 hidden layers. The circuit designs to carry out classification algorithms are described below.

2.1. Multiplication

The circuit block and timing diagram of a time-difference amplifier are shown in Fig. 2. The time difference, \( \Delta t_{\text{in}} \), between the input signals \( V_{\text{in1}} \) and \( V_{\text{in2}} \) is amplified through the delay propagation. A delay-locked loop (DLL) is used to reduce the sensitivity over process, voltage, and temperature variations, so that the output signals with the precise \( N \) times of input time difference can be generated.

There are two delay lines in the circuit, and each delay line contains \( N + 1 \) identical delay cells. The delays of cells in the constant delay line (CDL) are static during operation. However, the delays of cells in the voltage-controlled delay line (VCDL) are controlled by the \( V_{\text{CTRL}} \) signal generated by the feedback loop.

The input signals, \( V_{\text{in1}} \) and \( V_{\text{in2}} \), have the same clock periods, but with a time difference, \( \Delta t_{\text{in}} \). Output signals, \( V_{\text{out1,0}} \) and \( V_{\text{out2,0}} \), are connected to the phase/frequency detector (PFD), so the time difference between \( V_{\text{out1,0}} \) and \( V_{\text{out2,0}} \) are sensed by the PFD. UP/DN signals are generated according to the time difference and used in the charge pump (CP) to control currents for charging/discharging the loop filter. The resulted \( V_{\text{CTRL}} \) is used to modify the delay of VCDL to force \( V_{\text{out1,0}} \) to be in the same phase as \( V_{\text{out2,0}} \). As described above, the delay cell \( D_{\text{V1}} \) and \( D_{\text{V2}} \) are equally sized, so as \( D_{\text{C1}} \) and \( D_{\text{C2}} \). Therefore, the \( V_{\text{out1,1}} \) is one \( \Delta t_{\text{in}} \) ahead of \( V_{\text{out2,1}} \) instead of behind it. Then, \( V_{\text{out1,2}} \) is 2 times of \( \Delta t_{\text{in}} \) ahead of \( V_{\text{out2,2}} \), and \( V_{\text{out1,N}} \) is \( N \) times of \( \Delta t_{\text{in}} \) ahead of \( V_{\text{out2,N}} \).

To achieve higher resolution, the circuit is extended for more weight selections and a larger input range. \( M \) delay cells (\( D_{\text{A0}} - D_{\text{AM}} \) and \( D_{\text{B0}} - D_{\text{BM}} \)) are integrated in the delay-locked loop as shown in Fig. 3. The fine delay cells are used to divide the input time difference \( \Delta t_{\text{in}} \) to \( \Delta t_{\text{in}} / M \) as a unit delay that extends the input range by \( M \) times. Connecting \( V_{\text{out1,1-N}} \) and \( V_{\text{out2,1-N}} \) with two N-to-1 MUXs, the weights of \( \Delta t_{\text{in}} \) can be reconfigured to accomplish the multiplication. Therefore, the output time difference can be expressed as

\[ \Delta t_{\text{out}} = N \times \Delta t_{\text{in}}. \]  

(2)

Fig. 4 (a) illustrates the details of how the time difference between \( \text{CLK1} \) and \( \text{CLK2} \) is sensed through the PFD module. True-Single-Phase-Clock (TSPC)-based PFD is used for operating under low power supply

![Fig. 4. (a) The TSPC-base phase/frequency detector and (b) the charge pump. (b) Shows the circuit block diagram of the charge pump. After sensing UP and DN signals, the differential amplifiers charge or discharge the capacitor to change \( V_{\text{CTRL}} \). The charge pump adopts source-coupled pairs to steer currents, and cross-coupled pairs are used to increase the response time for low-voltage operations. The current of the voltage-controlled delay cell is controlled by \( V_{\text{CTRL}} \) to change the delay. Additional two inverters are used as buffers to shape the output signals for low power supply voltages.](fig4.png)
voltage [19]. While a rising edge on CLK 1 turns on M 5, the drain of M 5 is discharged so that DN goes high. In the same way, a rising edge on CLK 2 discharges the drain of M 11, so that UP goes high. Reset is triggered when both drain of M 5 and M11 go low to discharge the drain of M 3 and M9. It leads to forcing the drain of M 5 and M11 to go high. Therefore, if CLK 1 is ahead of CLK 2, the PFD sends out UP signal. If CLK 1 is behind of CLK 2, the DN signal is sent out.

2.2. Summation

To perform the inner-product operations, it requires the summation of several weighted time differences. Fig. 5 shows the presented time-domain inner-product architecture. Two DLLs are cascaded to sum up the weighted \( \Delta t_{in1} \) and \( \Delta t_{in2} \), where \( \Delta t_{in1} \) is the initial time difference between \( V_{in1_1} \) and \( V_{in1_2} \), and \( \Delta t_{in2} \) is initial time difference between \( V_{in2_1} \) and \( V_{in2_2} \). In both configurations of Stage_1 and Stage_2, the transition edge of \( V_{out1_1} \) is equal to \( V_{out1_2} \) and the transition edge of \( V_{out2_1} \) is equal to \( V_{out2_2} \) because VCDL1 and VCDL2 are adjusted by \( V_{CTRL1} \) and \( V_{CTRL2} \) in the feedback loop.

In order to propagate the weighted time differences from Stage_1 to Stage_2, the inputs of the delay cells outside the loop in Stage_2 are routed to the outputs of Stage_1, so the start points of the delay line VCDL2 and CDL2 are characterized by previous outputs of Stage 1. Therefore, the weighted delays are accumulated from different stages through the cascaded delay lines to acquire the sum shown in different colors in the figure. For example, if the SEL 1 is 3 and SEL 2 is 4, the time
difference between \( V_{O2,1} \) and \( V_{O2,2} \) is equal to \((3 \times \Delta t_{in1} + 4 \times \Delta t_{in2})\).

2.3. Pseudo-sigmoid activation function generator

The nonlinear activation functions are used in the hidden and output neurons to estimate the class probability for a given multiplication and accumulation result. A comparator shown in Fig. 6 has been designed as the pseudo-sigmoid activation function generator to transform the summation to probabilities. The summation of the weighted time differences controls charging time of the capacitors at the inputs of the regenerative comparator to perform logistic regression.

The regenerative sense amplifiers are usually used as comparators because the amplification is not required to be linear and achieves smaller delay time with positive feedback. The regenerative comparators can be simplified as the back-to-back inverter-based dynamic latch with its model shown in Fig. 7.

The output voltage can be calculated as

\[
V_{out}(t) = V_1(t) - V_2(t)
\]

(3)

where \( \tau = \sqrt{C_1 / Gm_1} \) and \( \alpha = \sqrt{C_2 / Gm_2} \). If the comparator is perfectly matched without any process variations, the output voltage of positive feedback characteristic of the dynamic latch can be expressed as

\[
V_{out}(t) = (V_{1,0} - V_{2,0}) \cdot e^{-t/\tau}
\]

(4)

The comparator output will regenerate more quickly with larger initial input difference as in Fig. 8. Therefore, the inverse of the exponential characteristic is utilized as the logistic sigmoid function

\[
f(x) = \frac{1}{1 + e^{-x}}
\]

(5)

3. Simulation results

The proposed circuits were simulated in a 65 nm CMOS process in Cadence and modeled as neuron cells for the system level simulation. Fig. 9 shows the simulation results of time-domain multiplication. The upper sub-figure shows that the delay between \( V_{Out1} \) and \( V_{Out2} \) is 0.100 \( \mu s \) when the initial input time difference is 0.1 \( \mu s \) and weighting of 1x is applied. The other 2 sub-figures show the weight setting of 4 and 8 and the corresponding time differences between the outputs. To compromise between speed and power consumption, the delay line is designed to carry out 16 times of delay multiplication. Fig. 10 shows the simulation results of 4-bit multiplication and accumulation that result in a summation of 5-bit matrix operation. The output time difference is changed linearly according to the weight values.

Fig. 11 shows the comparison of the normalized transfer curve of the presented activation function versus the standard sigmoid distribution. This pseudo-sigmoid logistic regression can be fitted as follows:

\[
f(x) = 0.9915 + \frac{1}{1 + e^{-1.18x}}
\]

(6)

The simulated transfer curve of the presented comparator demonstrated the s-shaped pseudo sigmoid function to transform inputs to probabilities. The resolution of the activation function is not limited by the digital levels because of its operation in time domain.

The system level demonstration was carried out in MATLAB for training and classification. Fig. 12 shows the training and testing setup. The time-domain classifier was trained with an off-chip engine. The system was evaluated by classifying the cardiac arrhythmia from the MIT-BIH arrhythmia database [20]. The ECG data that the experiments used is sampled at 360 Hz. Therefore, the classification results can be obtained with 20 k operations per second since the delay can be propagated to the next in the pipeline. The presented classifier achieves 90.5%
accuracy detection. Fig. 13 shows that power consumption scales with power supply voltages. Representing signals in time domain is not only beneficial from technology scaling, but also save significant power with lower power supply voltages. Unlike the conventional classification engines with data converters including ADCs and DACs to interface with the sensors, the time-domain operation can survive lower power supply voltages with lower operation speeds. The power supply can be even lowered to 0.4 V more complicated circuits/power consumption as shown in Ref. [21]. In this work the power supply of 0.9 V can be achieved without sacrificing operation speed to achieve best tradeoff between power consumption and operation speed. Since the calculated results are propagated to the next stage in the pipeline, the proposed architecture achieves 20 k operations per second per unit with power consumption of 240.34 nW. The estimated area for each neuron is less than 40 $\mu$m$^2$, which means less than 0.01 mm$^2$ is added for integration of the classifier into the sensor. Table 1 summarizes the performance and comparison with the other state-of-the-art works for in-sensor computing.

4. Conclusions

To eliminate the needs to continuously transmit complex signals to the aggregator for remote monitoring, a low-power time-domain in-sensor classifier that locally extracts critical features for rapid analysis is presented in this paper. The presented cascaded architecture utilizes DLLs to perform precise multiplication and accumulation. Through a pseudo-sigmoid activation function, the probability for the inner-product result is then estimated. Time-domain operations consume minimal energy under low supply voltages. Hence, the time-domain classifier can be
integrated with edge sensors to enable long-term continuous monitoring biomedical signals.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ethan Chen: Conceptualization, Methodology, Investigation, Software, Data curation, Visualization, Writing - original draft. Vanessa Chen: Conceptualization, Methodology, Supervision, Writing - review & editing, Project administration, Funding acquisition.

Acknowledgements

This work is supported by the National Science Foundation under Grant No. 1953801.

References


Table 1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>130 nm CMOS</td>
<td>130 nm CMOS</td>
<td>65 nm CMOS</td>
<td>65 nm CMOS</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>0.9–1.2 V</td>
<td>1.2 V</td>
<td>–</td>
<td>0.9 V</td>
</tr>
<tr>
<td>Operations per Second</td>
<td>1 k</td>
<td>20 k</td>
<td>100 k</td>
<td>20 k</td>
</tr>
<tr>
<td>Required Clock Frequency</td>
<td>2 kHz</td>
<td>20 kHz</td>
<td>100 MHz</td>
<td>20 kHz</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>28 nW</td>
<td>663.6 nW</td>
<td>243.7 μW</td>
<td>240.34 nW</td>
</tr>
<tr>
<td>Function</td>
<td>Compressive Sensing</td>
<td>Matrix-Multiplying</td>
<td>Matrix-Multiplying</td>
<td>Multiply-Accumulate &amp; Activation Function</td>
</tr>
</tbody>
</table>