Design of CMOS Quadratic Neural Networks

S. Mehdi Fakhraie, J. M. Xu, K. C. Smith

Department of Electrical and Computer Engineering,
University of Toronto,
Toronto, Ontario, Canada M5S 1A4

mehdi@vrg.utoronto.ca, xujm@eecg.utoronto.ca, kcsmith@csri.utoronto.ca

Abstract

In this work, we introduce an analog implementation technique for a class of quadratic artificial neural networks, in which the intrinsic quadratic characteristic of an MOS transistor in its saturation region has been employed. Test chips implemented in a 1.2 μm CMOS technology have shown full functionality as designed and demonstrate circular discriminating surfaces.

1. Neural Networks with Closed-Boundary Discriminating Surfaces

Research in artificial neural networks (ANNs) was originally inspired by an analogy with biological neural networks. However, the tremendous growth of "applied research" on neural-net applications has brought another strong driving force to the field. Requirements enforced by these new applications are pushing the researchers in a direction presumably not in full accordance with biological prototypes. At present, the bigger concern is to be more responsive to expanding applications by providing better and more efficient implementing means. This intention is further constrained by another motivation to provide fast, integrated, and efficient parallel implementations.

According to the most widely used model, a synapse multiplies its input by its weight value and provides a proportional output. In each neuron, outputs from different synapses are added together and then passed through a thresholding output function. This construct can constitute a linear discriminating surface in a pattern-classification context. In this type of architecture, vector-matrix multiplication is the calculation most frequently used. Fig. 1 shows a pictorial example for a 2-D input space. As shown, depending on the position of the input point, whether above, below or on the discriminating line implemented by the neuron, the output varies between its highest and lowest values.

However, in applications where several distinct clusters of data exist and one wants to isolate any of them from neighboring clusters, many linear discriminating surfaces are required. Correspondingly, providing a processing element with closed-boundary discriminating surfaces (in contrast to using unbounded hyper-planes) would be very resource-intensive.

Neurons with closed-boundary discriminating surfaces have already been introduced in the literature [1-4]. Their superiority to networks with linear synapses and linear discriminating surfaces has been demonstrated in certain applications including pattern recognition, character recognition, and data clustering where distinct classes of data exist. However, in a real-world application, there is still a major obstacle which must be removed before one can justify the use of these new characteristic functions as an alternative to conventional linear synapses, and that is the feasibility of a real-time, parallel hardware implementation using some existing electronics technology.

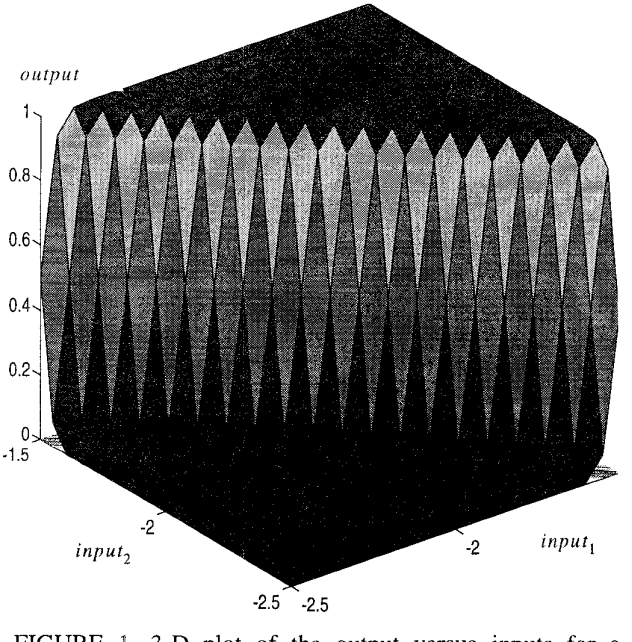

FIGURE 1. 3-D plot of the output versus inputs for a two-input neuron with linear synapses. The output is low, high, or zero depending on the position of the input point.

0-7803-2553-2/95/$4.00

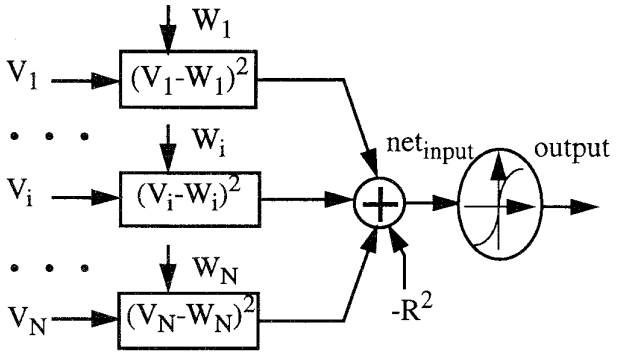

FIGURE 2. Block diagram of an analog quadratic neuron (AQN).

Accordingly, in this paper, we introduce a new implementation approach which better fits nonlinear pattern-recognition problems while employing intrinsic characteristics of MOS transistors operating in their saturation region. Further, while at first sight we seem to be departing from the most widely accepted model of an artificial neuron, such is not actually the case; for, considering the natural harmony of a biological neuron operating in conjunction with its surrounding environment, our processing elements are actually more neural-like, for they better match and employ properties of their natural existing environment, which is silicon!

2. Overview of Our Design

In work which we will now describe, we have selected an approach in which hyper-spheres are used as discriminating surfaces for our neurons. Fig. 2 shows a block diagram of the architecture we have implemented in this work. The weight vector $(w_1, \ldots, w_N)$ determines the centre of the discriminating hyper-sphere and the term $-R^2$ determines its radius. We can realize differently-sized discriminating boundaries by adjusting these parameters. The basic operations of the synaptic part of our quadratic neuron are calculating the difference of its two inputs and providing an output proportional to the square of that difference. In this way, the new difference-squaring operations are substituted for vector-matrix multiplication in feedforward perceptrons.

As well, this design will serve a dual purpose: it is that the collection of difference-squaring synapses in our design calculate the Euclidean distance between weight and input vectors, and thereby, can serve as the building block of an unsupervised-competitive-learning hardware. Another important application of our implementation scheme is in function approximation. As shown in Fig. 6, the output of an analog quadratic neuron (AQN) is a localized bump function with controllable radius. Therefore a single neuron of this type serves in place of $(2N+1)$ conventional neurons in the solution of N-dimensional function-approximation problems [7].

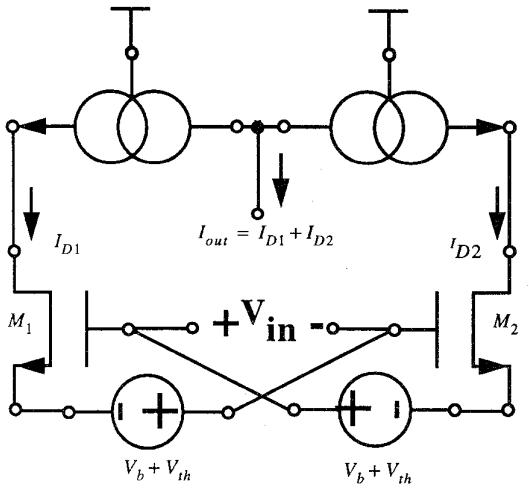

FIGURE 3. Basic diagram demonstrating the operation of a symmetric MOS subtracting- squaring circuit.

3. Circuit Design

Difference and squaring are the most widely used operations in AQNNs. Correspondingly, the well-known differential amplifier is able to provide us with a current proportional to the difference in the input voltages. Moreover, to perform the squaring operation, the characteristic equation of a MOS transistor in saturation is appealing. However, in order to introduce our desired variables appropriately under the squaring operation some manipulation is necessary. Among many different alternatives for implementing the squaring block, we have used a compact architecture originally introduced in [5] and improved subsequently in [6].

To gain some initial insight into the operation of the basic circuit employed, we draw a simplified version of it as shown in Fig. 3.

Hence, we assume M_1 and M_2 are matched and operating in their saturation region. Also we assume the channel length, oxide thickness and operating voltages are such that to a first-order approximation, the quadratic MOS characteristic equation applies as follows:

$$I_{D1} = K \cdot (V_{GS1} - V_{th})^2$$

$$I_{D2} = K \cdot (V_{GS2} - V_{th})^2 \qquad (1)$$

where $K = \frac{1}{2}\mu C_{OX}\frac{W}{L}$. Then, as seen in Fig. 3 we have:

$$I_{out} = I_{D1} + I_{D2} = K \cdot \left[(V_{GS1} - V_{th})^2 + (V_{GS2} - V_{th})^2\right] \qquad (2)$$

If we write two loop equations for the input loops, we conclude:

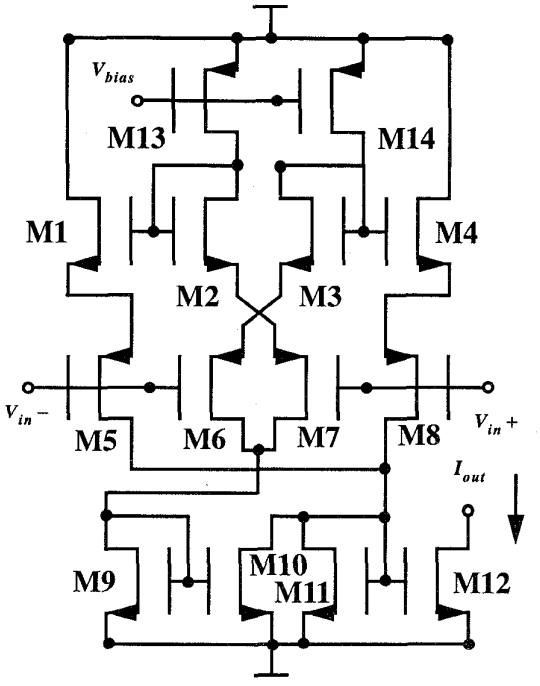

FIGURE 4. Schematic diagram of the subtracting-squaring circuit used as the basic synaptic unit.

$$-V_{in}+V_{GS1}-V_b-V_{th}=0\Rightarrow V_{GS1}=V_{in}+V_b+V_{th}$$

$$-V_{in}+V_{th}+V_b-V_{GS2}=0\Rightarrow V_{GS2}=-V_{in}+V_b+V_{th} \quad (3)$$

By substituting Eqs. 3 into Eq. 2, we get:

$$I_{out}=2K\cdot\left[V_b^2+V_{in}^2\right]=2K\cdot V_b^2+2K\cdot V_{in}^2 \quad (4)$$

Therefore, if $I_{sub}=2K\cdot V_b^2$ is subtracted from I_{out} we obtain:

$$I_{sum}=I_{out}-I_{sub}=2K\cdot V_{in}^2 \quad (5)$$

which is a current proportional to the square of the input differential voltage. This circuit has several drawbacks including current flow in the control path and requires implementation of floating equivalent voltage sources. However, these problems have been solved by replacing M_1 and M_2 with a "CMOS pair" as discussed in [6]. Also a pair of NMOS and PMOS transistors biased with a current source can be utilized to implement the required floating voltage sources. Applying these modifications to Fig. 3, and using the fact that I_{sub} in Eq. 5 can be obtained by adding the currents in the two floating-source branches, yields the final circuit used in our design, as shown in Fig. 4.

The output currents of the synaptic units are added together in a current-accumulator circuit, and a current proportional to R^2 is subtracted from them. This provides us with the equation $(X_1-W_1)^2+\ldots+(X_N-W_N)^2-R^2=u$. The sign of u determines whether the point $(X_1,\ldots,X_N)$ is inside $(u<0)$, or outside $(u>0)$, or on $(u=0)$ the discriminating hyper-sphere. Finally, a sigmoid output unit is employed which also performs the current-to-voltage conversion.

4. Hardware Implementation and Test Results:

Experimental CMOS chips have been fabricated using CMOS4S, the Northern Telecom 1.2 micron n-well double-metal double-polysilicon process provided through the Canadian Microelectronics Corporation (CMC). In this design, various building blocks have been fabricated in order that they be characterized individually. A complete two-synapse analog quadratic neuron demonstrating full circles as its discriminating surfaces has also been implemented.

A measurement setup has been prepared using an HP4145 semiconductor analyzer and 4 Keithley source and measurement units under control of a central computer. This setup enabled us to measure currents and voltages with μV and pA accuracy.

Subtracting-Squaring Unit (SSU): In the design of the subtracting-squaring unit (SSU), all of the transistors should operate in the saturation region in order to provide the functionality required. Therefore, appropriate sizing of transistors and choice of mirror topology is essential. Knowing the minimum gate length in the technology to be 1.2 μm, various gate lengths from 2 μm to 10 μm were tested. Cascode and single current sources were also evaluated. Comparison of different simulations together led us to conclude that using a single transistor as a current source with L= 3.4 μm and W/L = 10 provides a larger input dynamic range in comparison to using two cascoded transistors. Also it uses a smaller area and requires less supply-voltage drop. The circuit topology chosen is shown in Fig. 4. All of the transistors have $W=32\mu m$ and $L=3.2\mu m$. Each transistor is implemented as a parallel interconnection of many smaller units interleaved to realize matched couples.

Fig. 5 shows the test result for this unit. Here, the output current is measured while one of the inputs is fixed at $-1.8V$ and the other is swept from $-2.5V$ to $-1.1V$.

A Complete Neuron: A complete two-synapse AQN has been fabricated. It has been tested by scanning its inputs to extract its discriminating surfaces. We have tried to discriminate points inside and outside a circle with radius $R=0.5$ and centre at $(-1.8,-1.8)$. Therefore, one input at each synapse is fixed at $-1.8V$. In order to scan the whole workspace, the second input of one synapse is swept from $-2.5V$ to $-1.1V$ in $0.05V$ steps, while the other input of the second synapse is changed in $0.05V$ steps after each sweep. The resulting 3D plot of the input-output relation is as

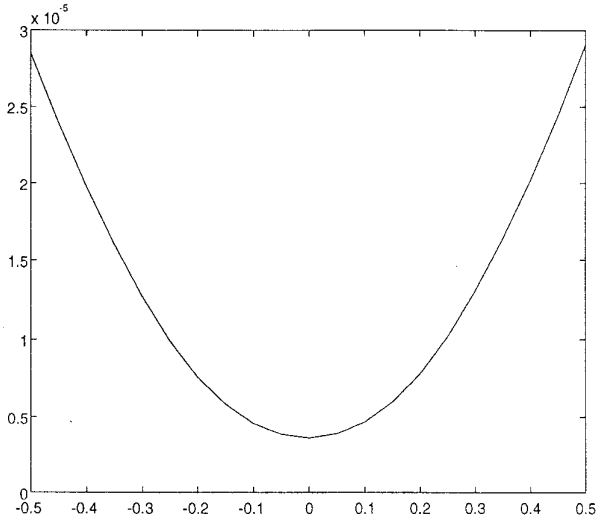

FIGURE 5. Test results for the subtracting-squaring unit. As seen, a symmetric quadratic output current results from a differential input voltage.

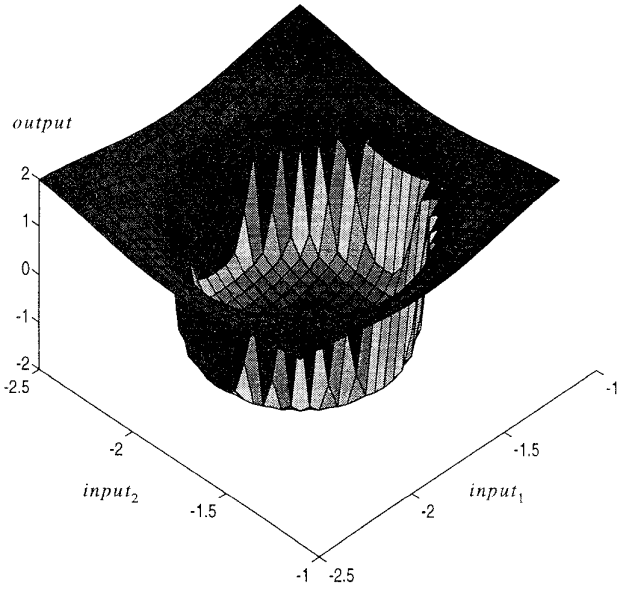

FIGURE 6. 3-D plot of measured output versus inputs of a two-synapse AQNN.

depicted in Fig. 6. Circular discriminating surfaces have been implemented in this two-quadratic-synapse neuron. In all, 841(29 x 29) input-output measurements have been used to construct this graph.

5. Conclusion

In this work, a direction for a class of application-driven hardware-compatible Artificial Neural Networks has been introduced. These networks demonstrate hyper-spherical closed-boundary discriminating surfaces when employed in a feedforward scheme. A simple MOS-compatible implementation exploiting inherent characteristics of MOS transistors in their saturation region has been shown. Sample chips has been fabricated in a CMOS $1.2\mu m$ process and full functionality has been verified by resulting measurements.

More information about design of the building blocks, chip fabrication, and measurements results are provided in [8].

References

[1] D. F. Specht, "Probabilistic Neural Networks," *Neural Networks*, vol. 3, pp. 109-118, 1990.

[2] J. Moody, and C. Darken, "Learning with localized receptive fields," in *Proc. 1988 Connectionist Models Summer School*, D. Touretzky, G. Hinton, and T. Sejnowski, eds., pp. 133-143, Morgan Kaufmann, 1988.

[3] D. L. Reilly, L. N. Cooper, C. Elbaum, "A Neural Network for Category Learning," *Biological Cybernetics*, vol. 45, pp. 35-41, 1982.

[4] D. J. Volper, S. E. Hampson, "Quadratic function nodes, use, structure, and training," *Neural Networks*, vol. 3, pp. 93-107, 1990.

[5] A. Nedungadi. and T. R. Viswanathan, "Design of linear CMOS transconductance elements," *IEEE Trans. Circuits and Systems*, vol. CAS-31, no 10, pp. 891-894, Oct. 1984.

[6] E. Seevinck, and R. F. Wassenaar, "A versatile CMOS linear transconductor / square-law function circuit," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 3, pp. 366-377, June 1987.

[7] A. Lapedes, and R. Farber, "How neural nets work," in *Neural Information Processing Systems*, D. Z. Anderson, ed., New York: American Institute of Physics, pp. 442-456, 1988.

[8] S. Mehdi Fakhraie, *VLSI-Compatible Implementations for Artificial Neural Networks*. Ph.D. dissertation, Department of electrical and Computer Engineering, University of Toronto, 1995.