ADAPTIVE RESONANCE THEORY NETWORKS USING INCREMENTAL COMMUNICATION

Ming Chen,* Ali A. Ghorbani, and Virendra C. Bhavsar†

Faculty of Computer Science
University of New Brunswick
Fredericton, NB, E3B 5A3, Canada
ghorbani@unb.ca bhavsar@unb.ca

Abstract

The incremental inter-node communication method is applied to the adaptive resonance theory 2 (ART2) networks. The incremental communication is aimed at reducing the communication costs of parallel and VLSI implementations of artificial neural networks. A node architecture incorporating the incremental communication is presented. A simulator is developed to study the behavior of ART2 networks with varying precisions of incremental data communication. Experiments are carried out to study the effects of the incremental communication on the convergence and savings in communication costs. We have found that even 7-bit precision in fixed-point and 13-bit (including 8-bit exponent) floating-point representations may be sufficient for the network to give the same results as those with conventional communication using 32-bit precision. The simulation results show that the limited precision errors are bounded and do not seriously affect the convergence of ART2 networks.

1 Introduction

The communication complexity of artificial neural networks is directly proportional to the number of inter-node connections. To reduce the cost of interconnection as well as intercommunication, we have proposed a new inter-node communication method named the incremental communication[6]. The effectiveness of the proposed communication scheme on multilayer feed-forward network architectures has been examined in the previous work [6, 7]. It has been shown through simulation that for some problems even 4-bit precision in fixed- and floating-point representations is sufficient for the network to converge. With 8-12 bit precisions almost the same results are obtained as those with the conventional communication using 32-bit precision [6]. The proposed method can lead to significant savings in the intercommunication cost for implementations of artificial neural networks on parallel computers as well as the interconnection cost of direct VLSI realizations. The method can be incorporated into most of the current learning algorithms in which inter-node communications are required. The main objective of this paper is to examine the effects of the limited precision incremental communication method on the convergence behavior of Adaptive Resonance Theory 2 (ART2) networks.

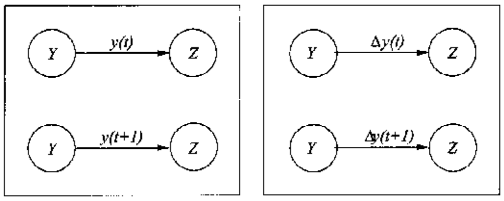
The paper is organized as follows. The following section introduces the concept of incremental communication method. Subsequently, the basic concepts of ART2 and the application of the incremental communication method to ART2 are given in Section 3. In Section 4, the simulation results for both fixed- and floating-point incremental communication methods are presented. Finally, conclusions are summarized.

*Presently at SWI Systemware Inc., Toronto, Ont., Canada.

†This research was partially funded by the grant OGP0089 from Natural Sciences and Engineering Research Council of Canada.

2 Incremental Communication

In ANNs an incremental value is defined as the amount of change in the input/output of a node at two consecutive steps n and n+1 [6].In incremental inter-node communication, instead of communicating the full magnitude of a variable, only the increment or decrement to its previous value is sent on a communication link. For example, assume that node Y has to communicate the signal y to node Z at different time instants, as shown in Figure 2. If $y(t)$ is the output of node Y at time t and $y(t+1)$ is its output at time $t+1$, in the conventional communication the communication link will carry the value $y(t+1)$. In contrast, in the incremental communication the communication link will carry the value $\Delta y(t+1)$, where $\Delta y(t+1) = y(t+1) - y(t)$. At the receiving end, the value $y(t+1)$ will be obtained by adding $\Delta y(t+1)$ to the previous value $y(t)$ stored at node Z.



a. Conventioal Communication

b. Incremental Communication

Figure 1: Conventional versus incremental communication.

The incremental value Δy can be represented in either fixed- or floating-point format.The fixed-point value may be represented in the integer or fractional form using a fewer number of bits (i.e., limited precision) than full-precision used for the signal y. In the floating-point representation, few bits of the mantissa and full value of the exponent are often used. When the incremental value Δy is limited to a smaller precision than the precision of y, we denote the incremental value by $\overline{\Delta} y$.

3 ART2 Networks

Adaptive Resonance Theory 2 (ART2) network is able to self-organize recognition categories for arbitrary sequenccs of binary or analog input patterns [1, 2].

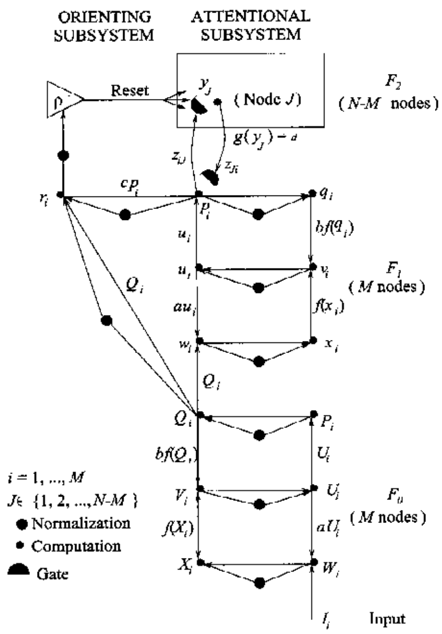


Figure 2: ART2 architecture

ART2 networks have three fields: a preprocessing field F_0, a layer of processing units called feature representation field F_1, and a layer of output units called category representation field F_2. F_1 and F_2 are fully connected in both directions via weighted connections called pathways. The set of pathways with corresponding weights is called an adaptive filter. The weights in the adaptive filter encode the long term memory (LTM) traces. The patterns of activation of F_1 and F_2 nodes are called short term memory (STM) traces. The connections leading from F_1 to F_2, and from F_2 to F_1 are called bottom-up and top-down adaptive filters, respectively. There are also corresponding bottom-up weights and top-down weights.

Figure 3 illustrates the ART2 architecture (adaptęd from [2]) used in our simulation studies. The F_0 field has the same structure as that of the F_1 field. The F_2 field consists of $N-M$ nodes that are fully connected. Since fields F_1 and F_2 are fully connected in both directions, ART2 network involves large amount of inter-node communication. Therefore, we propose the incremental communication method for the inter-node communication between the F_1 and F_2 fields.

We have incorporated the incremental communication method into the ART2 learning algorithm and modified the architecture of the F_0, F_1, and F_2 nodes using incremental communication. Figure 3 illustrates the F_1 node architecture using incremental communication method. In this figure, the incremental value $\overline{\Delta} z_{ji}(t+1)$ is input from a F_2 node whereas, incremental value $\overline{\Delta} u_i(t+1)$ is given out to other F_1 nodes. The function $\Phi(.)$ is used to truncate the value of $\Delta u_i(t+1)$ to limited precision value $\overline{\Delta} u_i(t+1)$. The α and β icons represent the points of connections to operators within the F_2 nodes. Note that the competitive learning part is omitted. The architecture of F_2 nodes is not given for the sake of brevity.

4 Simulation Studies

We investigate the effects of the limited precision fixed- and floating-point incremental values on the convergence of the network using continuous-valued weights. The conventional as well as incremental communication methods are implemented by modifying the simulator developed in [4]. In the conventional communication all the parameters are represented in full precision (32 bits), whereas in the incremental communication method all the incremental values are represented in reduced-precision. The incremental values of the parameters can be represented using either fixed- or floating-point schemes. With the fixed-point representation, the position of the binary point can be decided depending on the problem chosen.

Our experiments consist of training the ART2 network using the conventional and the incremental communication schemes while varying the precision of incremental values of STM (p_i and u_i, $i = 1, \ldots, M$) and top-down as well as bottom-up LTM in the fixed- and floating-point representations.

For ART2 network using conventional communication method, we have selected the architectures and parameters that resulted in good performance. To have comparable results, the same parameters and architectures are used with the incremental as well as the conventional communication methods. In the experiments for the fixed- and floating-point incremental communication, the precision of the incremental values is varied from 2 to 20 bits in the steps of one bit. For the fixed-point representation, we use 2 bits before the binary point and k bits after the binary point whereas, for the floating-point representation, 8-bit exponent and k-bit mantissa are used.

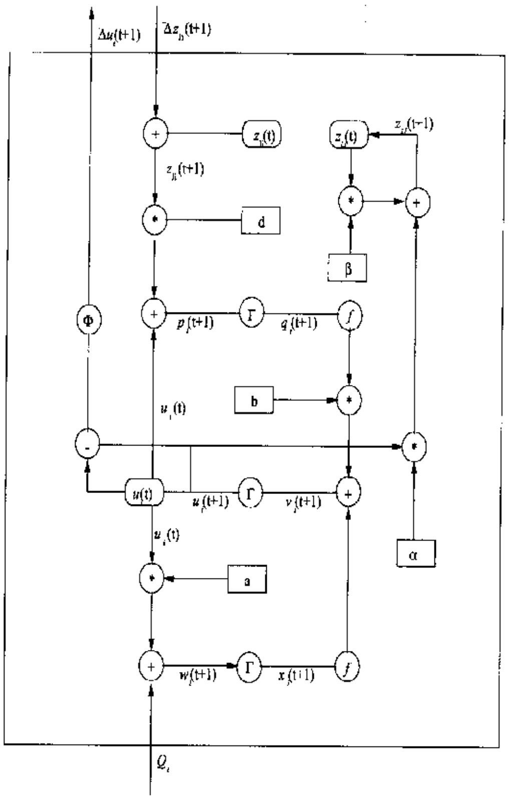


Figure 3: F1 node architecture

The test problem chosen for simulation studies is composed of 50 input patterns[3]. The learning goal is to classify 50 input patterns into different categories. The patterns that are grouped in the same category usually share some common features. We may get different number of categories by using different network parameters. We have carried out two sets of experiments (Experiments 1 and 2). For a set of experiments, we use the same network parameters and vary the type and precision of incremental values; different sets use different network parameters. The absolute average discrepancy between top-down expectations (the final top-down vectors for the established category after the network converges) of conventional and incremental communication is referred to as the *error* in the following subsections.

4.1 Experiment 1

For this experiment we set $\rho = 0.95$, $\theta = 0.23$, $a = b = 5$, $d = 0.8$, and $c = 0.22$ (see [3]for details). The training is considered complete when the category structure established on one complete presentation of the 50 inputs remains stable there after. When the network converges, the 50 input patterns are classified into 14 recognition categories. For floating-point representation, 5-bit mantissa is found to be enough to obtain the same results as those with the conventional communication method. For the fixed-point representation, 2 bits to the left and 6 bits to the right of the binary point are found to be enough to get the same results as those of the conventional communication.

Figure 1(a) represents the error as a function of the precision of fixed- and floating-point representations. It is seen that as the precision of incremental values increases, the error decreases quickly. The error with floating-point representation is consistently lower than that with the fixed-point representation. This is expected since the floating-point representation allows a large dynamic range of values.

Figure 1(b) depicts the total number of iterations required by the network to converge in conventional and incremental communications with various precisions. It is seen that the number of iterations required for the floating-point representation is always close to that for the conventional communication. When more than 8 bits are used for the fixed-point representation, the number of iterations is always close to, and sometimes even smaller than, that for the conventional (standard) communication.

The increase and decrease in computation communication costs for selected precisions are given in Table 1. In the table, t_{comm}, t_{comp1}, and t_{comp2} represent the increase/decrease in communication times, computation times in Experiment 1 and computation times in Experiment 2, respectively. Note that the negative values indicate the decrease, and the positive values indicate the increase in times. It is seen that the use of incremental communication results in substantial savings in communication time with very small, if at all, increase in computation time.

Table 1: The computation and communication times for Experiments 1 and 2.

Precision	Fixed-point			Floating point		
k	t_{comm}	t_{comp1}	t_{comp2}	t_{comm}	t_{comp1}	t_{comp2}
10	-62.5%	-0.39%	-1.2%	-43.7%	0%	-0.03%
8	-68.7%	-0.07%	-1.2%	-50.0%	-0.02%	0%
6	-75.0%	+5.59%	+0.8%	-56.2%	-0.23%	-0.29%
5	-78.1%	+3.81%	+4.4%	-59.3%	+0.07%	-0.09%

4.2 Experiment 2

For this experiment, we set $\rho = 0.98$, $\theta = 0.21$, $a = b = 5$, $d = 0.8$, and $c = 0.22$. The 50 input patterns used in Experiment 1 get classified into 21 categories due to higher value of vigilance (ρ) compared to Experiment 1. For both the floating-point and the fixed-point representations $k = 5$ gives the same results as those with the conventional communication method.

Figure 2 shows the error and the total number of iterations required by the network to converge in the conventional and the incremental communications with various precisions. Table 1 gives decrease/increase in the computation and communication times for different number representations. The behaviors observed in this experiment are very similar to those of the Experiment 1.

5 Conclusions

In this paper, we have examined the effects of the limited precision incremental communication method on the convergence behavior of Adaptive Resonance Theory 2 (ART2) networks. We have successfully incorporated the incremental communication method in the learning algorithm of the ART2 networks. Simulation results show that the increase in the number of iterations required for the convergence with incremental communication is found to be small. In fact, in some cases with an appropriate number of bits for incremental communication, the number of iterations is very close to and sometimes even smaller or the same as in conventional communication. It is shown that the communication costs are substantially reduced using the incremental communication method.

We have also carried out an error analysis to justify the behavior and explain the effects of limited precision incremental communication [3]. The analytical results are found to be in agreement with the simulation results.

In conclusion, the incremental communication method can considerably save communication times and interconnection costs. We have demonstrated that the incremental communication scheme is not only suitable for feedforward neural networks with supervised learning, but also for a class of recurrent neural networks with unsupervised learning.

References

[1] G. A. Carpenter and S. Grossberg, "ART2: self-organization of stable category recognition codes for analog input patterns," *Applied Optics*, vol. 26, pp. 4919-4930, 1987.

[2] G. A. Carpenter and S. Grossberg, "ART2-A: an adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, vol. 4, pp. 493-504, 1991.

[3] M. Chen, 'Incremental Communication for Adaptive Resonance Theory Networks,' MCS thesis, Faculty of Computer Science, University of New Brunswick, Fredericton, N.B., Canada, April 1998.

[4] P. Gaudiano, "ART2 - a simple ART2 simulation program", published in 1990; f*tp://cns-ftp.bu.edu/pub/ art2.shar.gz*, July 1997.

[5] A. A. Ghorbani and V. C. Bhavsar, "Artificial neural networks with incremental communication on parallel computers," in *Proc. of the 5th UNB Artificial Intelligence Symposium*, pp.15-28, Fredericton, N. B., Canada, August 1993.

[6] A. A. Ghorbani and V. C. Bhavsar, "Incremental communication for multilayer neural networks", *IEEE Trans. Neural Networks*, vol. 6, pp. 1375-1385, 1995.

[7] A. A. Ghorbani and V. C. Bhavsar, "Incremental communication for multilayer neural networks: Error analysis", *IEEE Trans. Neural Networks*, vol. 9, pp.68-82, 1998.

Figure 1: Experiment 1: (a) Error in top-down expectations, (b) The total number of iterations versus precision.

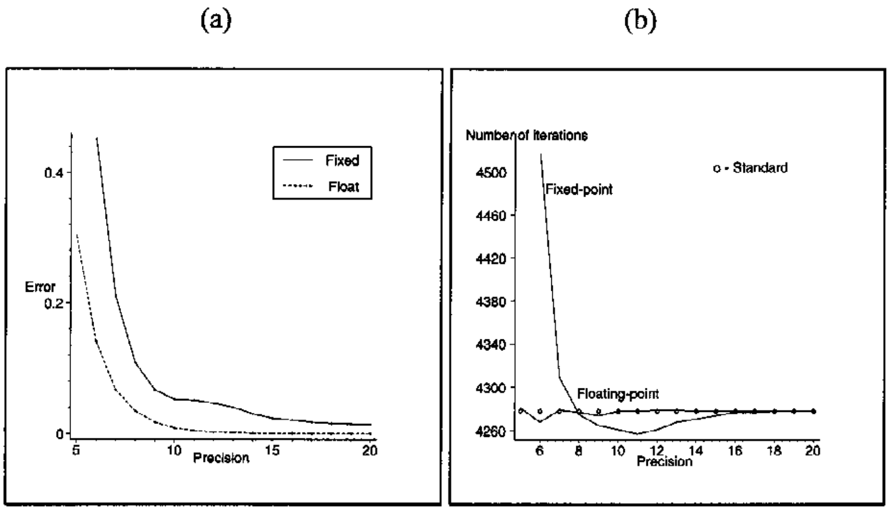
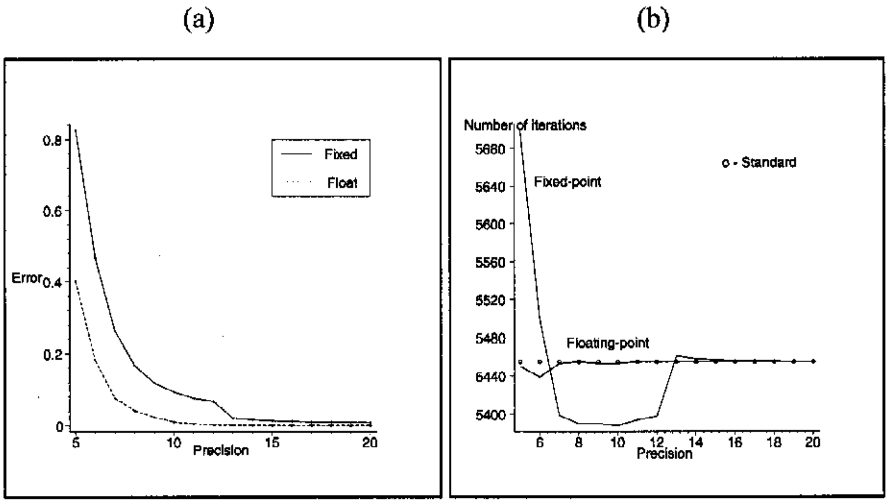


Figure 2: Experiment 2: (a) Error in top-down expectations, (b) The total number of iterations versus precision.


(a)
(b)
Fixed
Float
Error
Precision
Number of Iterations
o - Standard
Fixed-point
Floating-point
Precision