CASCADE ERROR PROJECTION WITH LOW BIT WEIGHT QUANTIZATION FOR HIGH ORDER CORRELATION DATA

Tuan A. Duong and Taher Daud
Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, CA 91109

Abstract: In this paper, we reinvestigate the solution for chaotic time series prediction problem using neural network approach. The nature of this problem is such that the data sequences are never repeated, but they are rather in chaotic region. However, these data sequences are correlated between past, present, and future data in high order. We use Cascade Error Projection (CEP) learning algorithm to capture the high order correlation between past and present data to predict a future data using limited weight quantization constraints. This will help to predict a future information that will provide us better estimation in time for intelligent control system. In our earlier work, it has been shown that CEP can sufficiently learn 5-8 bit parity problem with 4- or more bits, and color segmentation problem with 7- or more bits of weight quantization. In this paper, we demonstrate that chaotic time series can be learned and generalized well with as low as 4-bit weight quantization using round-off and truncation techniques. The results show that generalization feature will suffer less as more bit weight quantization is available and error surfaces with the round-off technique are more symmetric around zero than error surfaces with the truncation technique. This study suggests that CEP is an implementable learning technique for hardware consideration.

I-Introduction

There are many ill-defined problems in pattern recognition, classification, vision, and speech recognition that need to be solved in real time [1-3]. The solution by linear technique may not be suitable because its hyperplane solution may not be sufficient for very complex problems or it cannot provide a generalization feature for a new and unlearned data. Therefore, neural network is a good candidate to solve such problems. In addition, the neural network architecture, unlike sequential architecture, provides a massively parallel processing feature that offers tremendous speed only when implemented in hardware. From these evidences, neural network hardware is defined as our motivation in this paper.

In our earlier publications [4-6], it was shown that CEP is an efficient hardware learning algorithm. It only required 4-bit weight quantization to solve 4-8 bit parity problems and 7-bit weight quantization to reproduce the same accuracy results as 64-bit weight quantization with two more hidden units added for color segmentation problem[7]. We now broaden the application of CEP to solve the chaotic Mackey-Glass time series prediction problem.

The nature of this problem is such that data never repeats itself, but is rather chaotic. To capture this prediction, the transformation for the future data must contain high order

correlation of the past and present data. This study will, once again, help us to confirm the potential of CEP to such high order correlation for prediction which may be useful for intelligent control or robust data validation.

II. ARCHITECTURE OF CEP:

The CEP architecture is shown in Figure 1 where X is the input and O is the output set. Hidden units are added one at a time as needed.

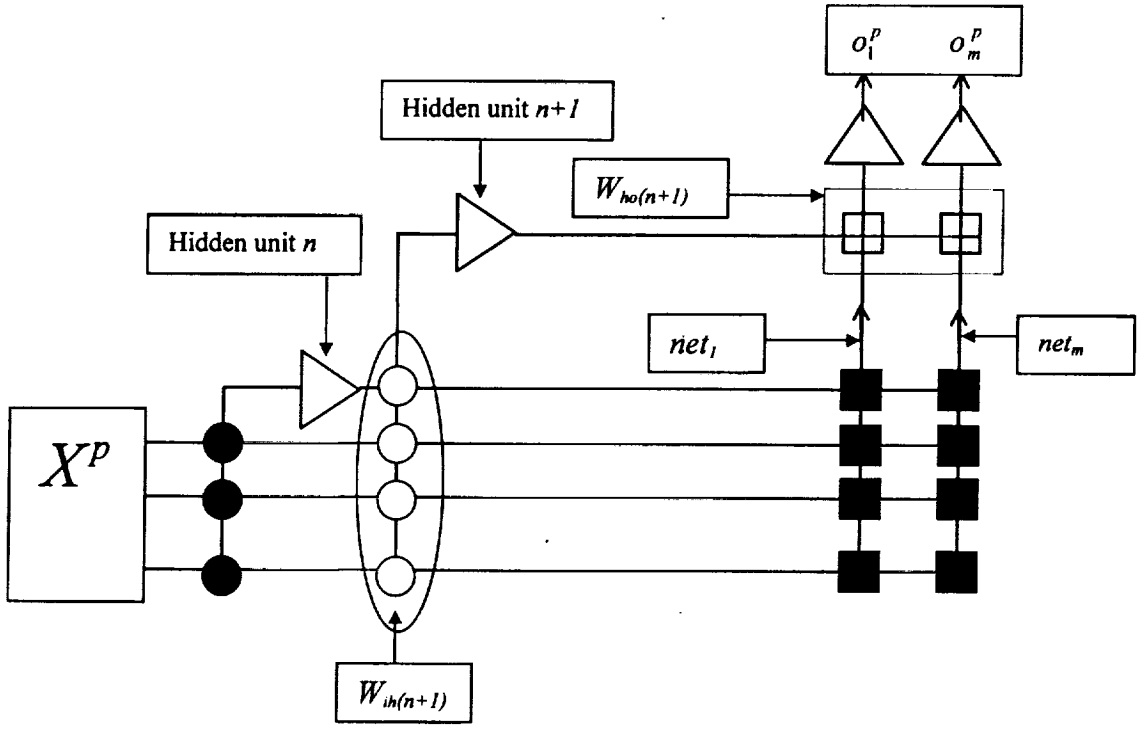

Figure 1: Assume that there are (n+1) hidden units in the network and the blank squares and circles are the weight components which determine the weight values by learning or calculating.

$\varepsilon_o^p = t_o^p - o_o^p(n)$ denotes the error between output element o and training pattern p with target t and actual output $o(n)$ where n indicates that the output has n hidden units in the network;

$f'^p_o(n)$ denotes the output transfer function derivative with respect to the input with the index of the output o (o={1,m}) and the training pattern p;

$f_h^p(n+1)$ denotes the function of hidden unit n+1 and training pattern p;

X^p denotes the input pattern p; and $|X|$ denotes the Euclidean length of vector X.

it has been shown [5-6] that:

$$\Delta E = \sum_{p=1}^{P} \sum_{o=1}^{m} \{\varepsilon_o^p f'^p_o f_h^p(n+1)\} \qquad (1)$$

Our goal is to maximize a difference energy function ΔE with respect to W_{ih} through training. Then, we can calculate W_{ho} which can be determined by equation below:

$$W_{ho} = \frac{\sum_{p=1}^{P} \varepsilon_o^p f'^p_o f_h^p(n+1)}{\sum_{p=1}^{P} [f'^p_o f_h^p(n+1)]^2} \qquad (2)$$

The details of CEP procedure can be found in [4].

Differences between CEP and CC:

The common goal of CEP and Cascade Correlation (CC) [8] is to maximize ΔE through weight set W_{ih}. However, the following differences can be noted:

1. The technique to achieve the maximization of ΔE with CEP is based on perceptron learning; versus covariance/correlation learning for CC. From the hardware view point, perceptron learning using stochastic technique is easier to implement in hardware as compared to covariance/correlation, using a batch technique.
2. CEP uses one hidden unit at a time with zero initial weight while CC uses a pool of candidate hidden units with different random initial weights for new hidden unit and picks the best candidate out of this pool.
3. $W_{ho}(t+1)$ is the only component needed to be calculated in CEP whereas in CC, $W_{ho}(j)$ with j=1:t+1 and W_{io} are both relearned, when a new hidden unit is added.
4. Most important is that equations (1) and (2) are obtained in CEP through a mathematical analysis whereas equation (1) is empirically introduced in CC.

From above, the weight sets ($W_{ho}(n+1)$ and $W_{ih}(n+1)$) which relate to a new hidden unit n+1 are the only sets to be learned in CEP. From this strategy, the algorithm is able to manipulate the dynamical stepsize of weight discretization to be proportional to the previous energy to achieve the efficient limited weight quantization. The results of this technique were analyzed and were published elsewhere [4, 9].

III. SIMULATION:

Problem:

The chaotic time series can be defined as follows [10]:

$$\dot{x}(t) = -bx(t) + \frac{ax(t-\tau)}{1+x^{10}(t-\tau)}$$

With a=0.2, b=0.1, and τ=17

For chaotic time series prediction problem, the input to the network is x_i, x_{i+1}, x_{i+2}, and x_{i+3} and the corresponding target is x_{i+4}. The number of training set values is 351 and the number of test data values is 651. We trained this data set with different bit weight

quantization varying from 4- to 6-bit and floating point machine weight (64-bit for double precision) with two techniques: round-off and truncation, to quantize ΔW.

Simulation Results:

Training performance:

For the training phase, the results are summarized in Table 1. In this table, we only present 3 hidden units to be added for this study. For this first block, we used floating point (64-bit) to train the 351 data. The learning performs well with one hidden unit and it continues to improve when more hidden units are added as shown by the root mean square (RMS) and standard deviation (STD) values.
With 6- and 4-bit weight quantization, the learning performance with the two methods of weight quantization are very close; however, the analysis [4] suggested that the round-off would perform better in learning. In 5-8 bit parity problems, the simulation results agreed with the theoretical analysis [4].

N# of hidden unit	1	2	3
64-bit floating point machine			
RMS	0.019207	0.013286	0.006507
STD	0.094256	0.006506	0.032055
Round-off technique			
6-bit Weight Quantization			
RMS	0.024143	0.015591	0.014715
STD	0.214052	0.206704	0.166339
4-bit Weight Quantization			
RMS	0.035043	0.019509	0.011999
STD	0.105488	0.245165	0.028074
Truncation Technique			
6-bit Weight Quantization			
RMS	0.025086	0.018955	0.012648
STD	0.141246	0.115860	0.121376
4-bit Weight Quantization			
RMS	0.097276	0.020337	0.019972
STD	1.512927	0.303675	0.292858

Table 1: The training performance of CEP for the chaotic Mackey-Glass time series prediction problem using the round-off and truncation techniques for weight quantization.

Generalization performance:

After completion of the training phase, our network was set up to test a set of unlearned data that contained 651 test values. The test results of the network performance with the round-off and the truncation methods for weight quantization are given in Figures 2 and 3 respectively.

In Figure 2a and 3a plots, the prediction values with 64-bit floating point (double precision) as well as with 4- and 6-bit weight precision, are plotted along with calculated results. For clarity, the comparative errors are plotted in Figure 2b and 3b. As expected, the least errors (~1%) are with the 64-bit precision. However, with 4- and 6-bit weight precision, the errors are of the order of 2.5%, concentrated mainly at the sharp peaks and valleys.

Further, a comparison of results of Figures 2 and 3 show that the round-off method for weight quantization seems to work slightly better than the truncation method as by shown by the dotted curves in the two figures. Specially, it may be noted that the errors with the truncation method are more skewed below zero.

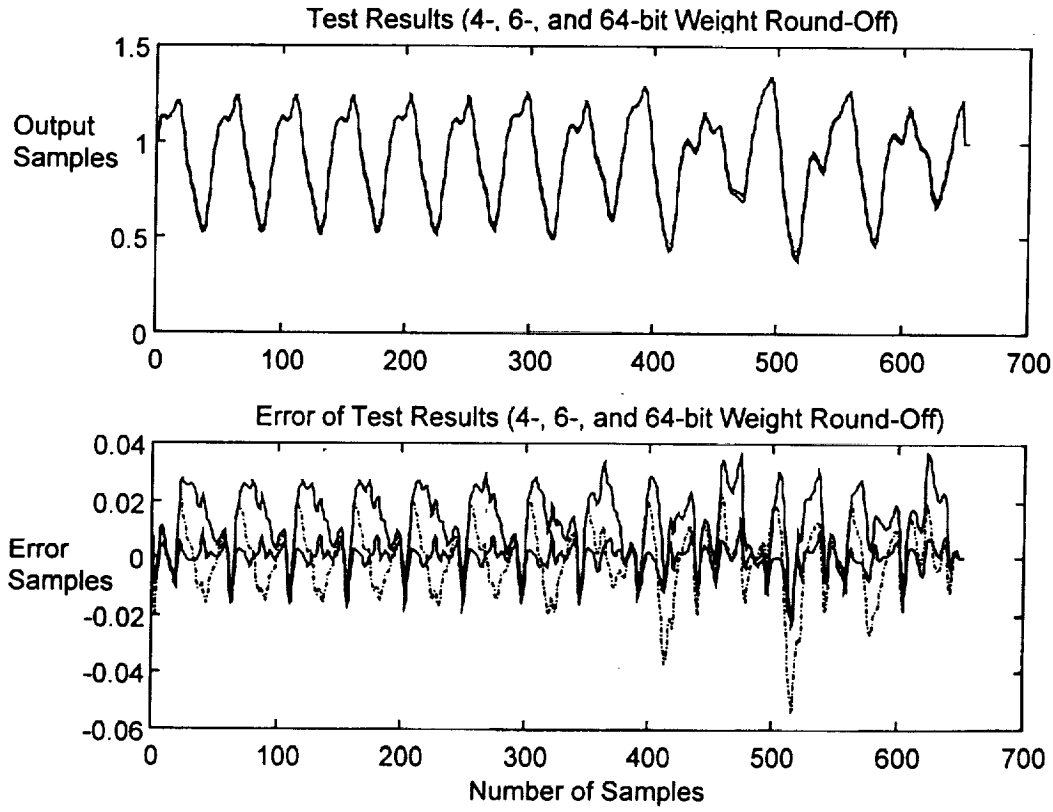

Figure 2: CEP prediction results with 4-, 6-, and 64-bit round-off weight quantization. The top trace (2a): prediction values; and the bottom trace (2b): the errors between prediction values and the target output which was generated from the chaotic equation above.

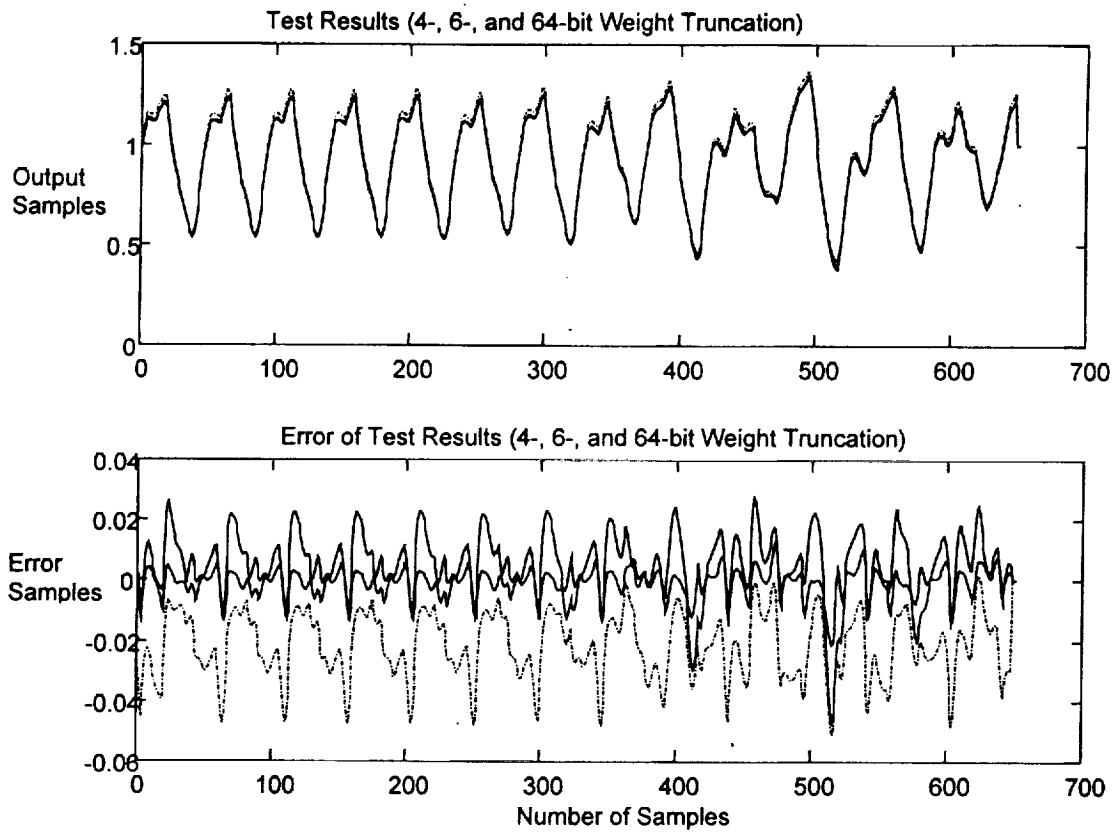

Figure 3: CEP prediction results with 4-, 6-, and 64-bit truncation for weight quantization. The top trace (3a): prediction values; and the bottom trace (3b): the errors between prediction values and the target output which was generated from the chaotic equation above.

Discussions:

CC[11] was reported under limited weight quantization for the 6-bit parity problem. The applied technique is very complicated, not suitable for hardware consideration, and required 8-bit weight resolution. In comparison, the CEP architecture with its learning algorithm could learn 6-bit parity problem with more than 3-bit weight resolution with ~100% accuracy [5]. For the chaotic time series prediction problem, we are not aware of any other results using limited weight quantization, specially down to only 4-bit. These results clearly demonstrate the power of CEP technique, specially as applied to the hardware implementation for taking advantage of the parallel processing and hence potential of very high speed of learning. This would lead to solution for complex control problems in real time when on-chip learning will be embedded in the silicon chip.

IV. CONCLUSIONS:

The advantages of the CEP learning algorithm can be summarized as follows:

- Simple perceptron learning procedure is applied.
- Learning scheme is tolerant of lower weight resolutions.

- A reliable model in learning neural networks as shown by the solutions of the benchmark problems.

Hence, CEP is a hardware implementable learning technique.

Acknowledgments:

The research described herein was performed by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology and was jointly sponsored by the Ballistic Missile Defense Organization/Innovative Science and Technology Office (BMDO/IST), and the National Aeronautics and Space Administration (NASA). The authors would like to thank Drs A. Stubberud and A. Thakoor for useful discussions.

References:

[1] T. A. Duong, T. Brown, M. Tran, H. Langenbacher, and T. Daud, "Analog VLSI neural network building block chips for hardware-in-the-loop learning," *Proc. IEEE/INNS Int'l Join Conf. on Neural Networks*, Beijing, China, Nov. 3-6, 1992.

[2] T. A. Duong et. al, "Low Power Analog Neurosynapse Chips for a 3-D "Sugarcube" Neuroprocessor," *Proc. of IEEE Intl' Conf. on Neural Networks*(ICNN/WCCI, June 28-July 2, 1994, Orlando, Florida), Vol III, pp. 1907-1911.

[3] B.E. Boser, E. Sackinger, J. Bromley, Y. LeCun, and L.D. Jackel, "An Analog Neural Network Processor with Programmable Topology," *IEEE Journal of Solid State Circuits*, vol. 26, NO. 12, Dec. 1991.

[4] T. A. Duong, *Cascade Error Projection-an efficient hardware learning theory, Ph.D. Thesis, UCI*, 1995.

[5] T.A. Duong, "*Cascade Error Projection-an efficient hardware learning algorithm,*" Proceeding Int'l IEEE/ICNN in Perth, Western Australia, Oct. 27-Dec 1, 1995, vol. 1, pp. 175-178 *(Invited Paper).*

[6] T.A. Duong, A. Stubberud, T. Daud, and A. Thakoor, "*Cascade Error Projection-A New Learning Algorithm,*" Proceeding Int'l IEEE/ICNN in Washington D.C., Jun. 3-Jun 7, 1996, vol. 1, pp. 229-234.

[7] E. Fiesler, L. Kempem, S. Campbell, T. Jannson, and T. A. Duong, "*Fuzzy neural chips for RGB sensor applications*, Accepted to SPIE SanDiego, July, 1998 *(Invited Paper).*

[8] S. E. Fahlmann, C. Lebiere, "The Cascade Correlation learning architecture," in *Advances in Neural Information Processing Systems II*, Ed: D. Touretzky, Morgan Kaufmann, San Mateo, CA, 1990, pp. 524-532.

[9] T.A. Duong, S.P. Eberhardt, T. Daud, and A. Thakoor, "*Learning in neural networks: VLSI implementation strategies,*" In: *Fuzzy logic and Neural Network Handbook*, Chap. 27, Ed: C.H. Chen, McGraw-Hill, 1996 .

[10] M. Mackey and L. Glass, "*Oscillations and chaos in physiological control systems,*" Science 197, pp. 287-289, 1977.

[11] M. Hoehfeld and S. Fahlman, "Learning with limited numerical precision using the cascade-correlation algorithm," *IEEE Trans. Neural Networks*, vol.3, No. 4, pp 602-611, July 1992.