A method of Constructing Light-Weight Circuit for Echo State Networks on Embedded systems

Makoto 'IT'oh (1310752)

School of Information Science,
Japan Advanced Institute of Science and Technology

August 5, 2017

Keywords: Echo State Networks, Circuit optimization, Embededd systems, Artificial Neural Network, motor control.

1 Introduction

Currently, in the field of ANN, there are many researches of computer vision that utilize CNN. In the field of ANN(motor control), ESN is a representative, which is a model of human cerebellum. And in general, a GPU is used for speeding up ANN process, and ESN can also be accelerated. But in the resource constraints environment like a small embedded system, it's difficult to place a GPU. Of course, there is also a small GPU, but sometimes it has unnecessary high precision for the calculation of ESN.

So this study proposes how to construct a light weight circuit with the necessary minimum precision and structure for the purpose of introducing ESN into a small embedded system. In particular, the study suggests a method of constructing light weight circuit that keeps an accuracy of ESN output wave, by analyzing a calculation accuracy of all nodes that are randomly connected each other in ESN. At first, the study defines ESN calculation and our proposing method. Second, it searches minimum hardware parameters by our proposing methods. Then, it describes a design of ESN circuits under minimum and original(not light weight) hardware parameters by verilog HDL, and evaluates these size and runtime.

2 Related work

Related works are as follow:

- About simplification of activating functions of CNN. [1]
- About selective usage of intermediate calculation results in CNN. [2]
- About minimum arithmetic bit-width of CNN that keeps perfomance. [3]
- About minimum complexity echo state network. [4]

- About simplification circuit of tanh(x). [5]

Standard ESN also uses the same sigmoid function as standard CNN, but there's no ESN report about the simplification of this function. And the selective usage of intermediate results is a good idea because ESN also has redundant intermediate values in its internal nodes, so it can reduce nodes and arithmetic bit-width, but there's no report about this point of view. Although there is a report about simplification of the hyperbolic tangent sigmoid function circuit, the accuracy is not sufficient for this study, so we ensured accuracy using mantissa bits.

While keeping up with the results of these previous report, this study focuses on "an accuracy bias of each neuron node in calculation of ESN" which has not been discussed in the previous ones and succeeds in reducing the circuit size.

3 Proposed Method

We propose four methods to reduce ESN hardware parameters.

- ESN node reduction by Sparsing Rate
- ESN node reduction by Graph Analysis Argorithm
- bit-width reduction of arithmetric circuit
- smplification of activating functions
- high/low accuracy calculation

1st method is to set the inside of ESN connection matrix element to 0 with the ratio of Sparsing Rate. (The order of setting 0 is ascending in this matrix's elements.) After ESN node reducing by Sparsing Rate, some of connection matrix elements are 0 in some places, then Graph Analysis Algorithm can be applied. 2nd method, Graph Analysis Algorithm, is using breadth-first search algorithm, and detects nodes that are not connected to input nodes. Since "a connection matrix element is 0" means "the two edge nodes are not connected in that element", and by this infomation, a breadth-first search algorithm can detect internal nodes which are not connected with input nodes. Then disconnected node paths (some of connection matrix's elements) are set to 0. 3rd method is making a filter that reduces the bit width in software arithmetic operation, and it is used for reducing the bits of the adders and multipliers conforming to IEEE 754 which are constructed in this study. 4th method is simplification of the activation function. Focusing on the fact that tanh can be approximated separately for linear and nonlinear sections, we designed a lookup table under that point of view. 5th method is basis on a hypothesis: "an accuracy bias of each neuron node in calculation of ESN". ESN nodes are randomly connected each other, and has state in each nodes. And each ESN node echoes its state value to neighbor nodes via connected paths. So if a node has many connections from neighbor, it needs high accuracy calculation. On the other hand, if a node has a few connections from neighbor, low accuracy calculation is sufficient. Basis on a hypothesis, 5th method selects high accuracy nodes in ESN. Selection criteria is a ranking of similarity between a node output waveform and model waveform.

4 Evaluation

In order to evaluate the effectiveness of the proposed method, we measured the accuracy of prediction wave using three model signals of learning object, MGt 17, Lorentz, and Rossler. First, we measured an accuracy of standard ESN's prediction wave under these condition:

- arithmetic bit width: all 16 bits
- a number of input nodes: 2
- a number of internal nodes: 1000
- a number of output nodes: 1

Next, in the ESN calculation on the software, we used the parameters of the proposed method while changing them. And we have searched for HW parameters where the predicted waveform accuracy of each model signal in the ESN after applying the method is about the same degree (about 98 percentage) as the predicted waveform accuracy of the standard ESN. Changing parameters are as follow:

- Sparsing Rate: 0.0 - 99.9
- use Graph Analysis Algorithm: yes or no
- arithmetic bit width(high, low): 2 - 16
- apply smplification of activating functions: yes or no
- high accuracy nodes: 1, 3, 5, 20, 50, 100

As a result of the search, the accuracy of the predicted waveform did not degrade even with the following parameters:

- Sparsing Rate: 99.9
- use Graph Analysis Algorithm: yes
- arithmetic bit width(high, low): 6 - 10
- apply smplification of activate function: yes
- high accuracy nodes: 1

The connection matrix was about 10^{-5} times as large. Finally, we designed the circuit before and after applying the proposed method, and described the code in HDL. In a design before applying the proposed method, we made a unit that contains MAC with small RAM. Since the size of "before" circuit may be large, we determined a degree of parallelism of the circuit after making this unit. Then we compared these 2 circuits, and the circuit size was about 1/100 and the execution time was about 1/250.

5 Conclusion

The hypothesis: “an accuracy bias of each neuron node in calculation of ESN”, was correct. So we succeeded in building light weight circuit by reducing redundant ESN nodes and arithmetic bit width. In addition, it is possible to realize more high-speed operation and implementation on a lower cost circuit. Finally, ESN is a model of human cerebellum, so the study made it possible to introduce a part of human motor control to robots.

References

[1] DAHL, George E.; SAINATH, Tara N.; HINTON, Geoffrey E. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013. p. 8609-8613.

[2] Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." Journal of Machine Learning Research 15.1 (2014): 1929-1958.

[3] Gupta, Suyog, et al. "Deep learning with limited numerical precision." CoRR, abs/1502.02551 392 (2015).

[4] Rodan, Ali, and Peter Tino. "Minimum complexity echo state network." IEEE transactions on neural networks 22.1 (2011): 131-144.

[5] Leboeuf, Karl, et al. "High speed VLSI implementation of the hyperbolic tangent sigmoid function." Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on. Vol. 1. IEEE, 2008.