Stochastic Computing based AI System for Mobile Devices

Su Yeon Jang, Student *Member, IEEE*, Young Hyun Yoon, Student *Member, IEEE*, and
Seung Eun Lee*, *Member, IEEE*
Dept, of Electronic Engineering, Seoul National University of Science and Technology, Seoul, Korea
*seung.lee@seoultech.ac.kr

***Abstract*-- In this paper, we present a stochastic computing based AI system for mobile devices. As technology in AI advances, more complex computations are required. In case of mobile devices, it is hard to accommodate the entire computations due to power and area limitations of an embedded system. As a stochastic computing replaces the complex computations into simple computations, the mobile devices are available to include the AI system. In order to verify our design, the embedded AI system including stochastic computing is implemented on an field-programmable gate array (FPGA), and we successfully demonstrated the feasibility of the proposal.**

I. Introduction

A technology advance across an artificial intelligence (AI) increases the computation complexity. The complex computations help the AI system to achieve high accuracy, but also consume a large number of power and area. Especially, the embedded system has limitations about area and power due to the characteristic of embedded devices [1]. For that reason, the embedded system is hard to contain many multiplications for computation in a restricted environment. In order to overcome the limitation, we need a specific computation module to reduce logic area and power. Stochastic computing is a low-complexity computation module. The stochastic computing consists of a bit-stream to express probability in a ratio of 0 and 1 [2]. The stochastic bit-stream represents the number of 1 as the size of the data. The stochastic computing performs multiplication with logic operations to reduce computation size. According to the smaller computation logic size, the system requires less power consumption. For that reason, the stochastic computing is available to be used in various fields to conduct many operations such as image processing, filter, and machine learning algorithm [3-5]. When the AI system includes the stochastic computing, the system is minimized than including the multiplications without stochastic computing [6]. Therefore, the stochastic computing is effective to the embedded AI system in terms of the area and power consumption.

In this paper, we propose a stochastic computing based AI system for mobile devices. The proposed stochastic computing performs the simple AND operation instead of the complex square operation. In order to verify the functionality of the stochastic computing, we demonstrated our design with the embedded AI system by implementing on an field-programmable gate array (FPGA).

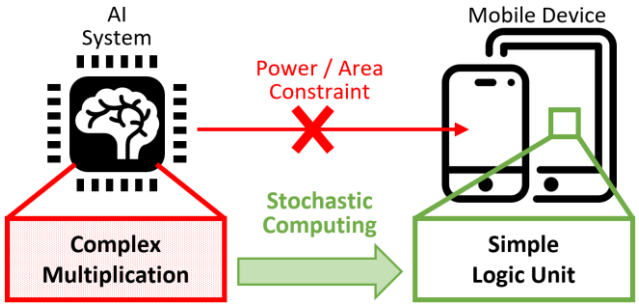


Fig. 1. Stochastic computing for implementing AI system on mobile devices

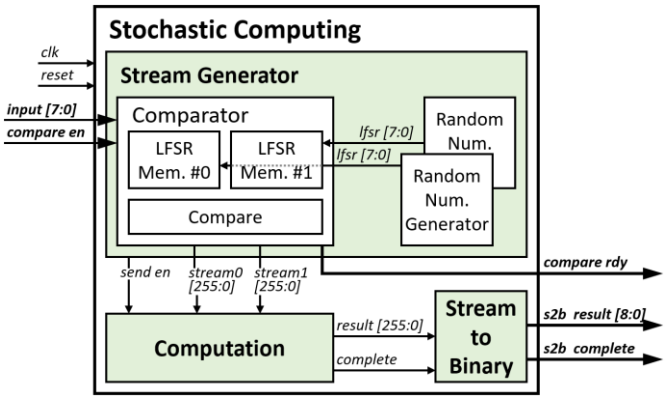


Fig. 2. Architecture of stochastic computing

II. Stochastic Computing

Fig. 2 shows the architecture of the stochastic computing. The stochastic computing consists of stream generator, computation, and stream to binary logics. By the characteristic of the stochastic computing, the stochastic computing needs to convert binary data into a stochastic bit-stream. The stream generator, which focuses to express the value of data stochastically, consists of random number generators and comparator. The random number generators use linear feedback shift registers (LFSRs) for hardware design and transfer the 8-bit LFSRs to the comparator. The comparator consists of dual LFSR memories and compare unit to generate bit-streams that the length is 2 power 8. Each LFSR memory has 8 widths and 128 depth to save the 8-bit LFSR from the random number generators. In order to generate the bit-stream, the input data is compared with both LFSR memories as the comparator receives input data. When the value of data is larger than the LFSR, the one bit of the bit-stream sets 1. After comparing between the data and the LFSRs, the computation logic transmits the generated bit-streams. The computation

978-1-7281-5186-1/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 25,2020 at 20:29:13 UTC from IEEE Xplore. Restrictions apply.

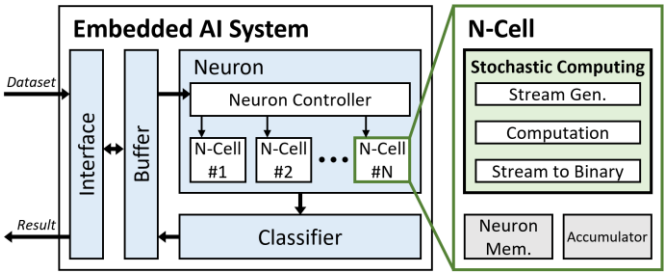


Fig. 3. Block diagram of the embedded AI system

logic operates the bit-streams by AND operation to substitute for multiplications when the send signal sets high state. After that, the computation result is transferred to the stream to binary logic to convert to binary data. When the stream to binary logic receives the complete signal from the computation logic, the stream to binary logic counts 1 in the operated bit-stream from the computation logic to convert from the bit-stream to the value of binary data.

III. Implementation and Result

In order to verify the functionality of our proposal, we design the embedded AI system including the stochastic computing. Fig. 3 shows the block diagram of the entire AI system. The system consists of interface, buffer, neuron, and classifier modules, and the neuron which enables the system learning and recognizing contains the stochastic computing module. For learning sequence, a train dataset is transmitted to the interface moudle. The dataset is delivered to the neuron controller in the neuron module through the buffer, and the controller sends the dataset to the pointed neuron cell. After the neuron cell finishes learning the dataset, the controller points the next neuron cell for learning and gets ready to be recognized. In order to recognize, the test dataset is transmitted to the neuron controller through the interface and buffer. For calculating the difference from the learned and test data in each neuron cell, the dataset is get into all of learned neuron cells and computed in simultaneously. The differences are calculated based on the Euclidean distance, and the multiply operations are substituted to the stochastic computing as the stochastic computing module is included in each neuron cell. The stochastic computing module performs the square operation and the result value is accumulated in the neuron cell. The accumulated values are transferred to the classifier module, and the classifier module based on the distance comparing algorithm finds out the nearest distance for recognition. Thereafter, the classified result is informed through the buffer and interface modules.

As the embedded AI system including stochastic computing is available to classify any dataset, we experimented with utilizing the system for recognizing the face expression. In Fig. 4, the experimental environment includes a camera in the host PC and an FPGA development board connected with the serial interface. The host PC receives a face image and pre-processes the image into the dataset for learning or recognizing. The FPGA, which is configured to the embedded AI system with the designed stochastic computing, receives the dataset and

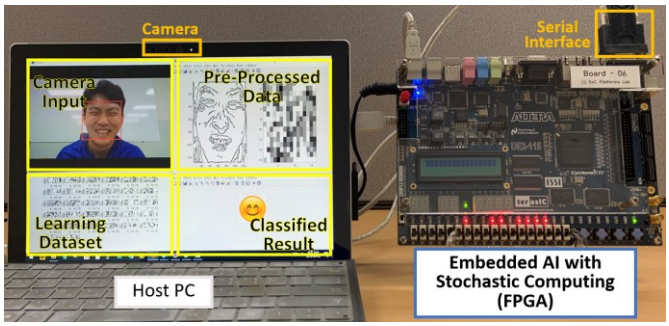


Fig. 4. Experimental environment and result

performs learning and recognizing. Through the classification with the system, the proper result is displayed on the host PC with claassified emoticon. Consequently, we verified the feasibility of the proposed stochastic computing successfully.

IV. Conclusion

In this paper, we propose the stochastic computing based AI system for mobile devices. In the embedded system which requires complex operations, the stochastic computing helps to replace the operations simple. Also, the stochastic computing enables to reduce the area and power consumption. By adopting the stochastic computing in the AI processor for face recognition, we implemented the proposed design on the FPGA. After demonstrating the functionality of the stochastic computing, we verified the feasibility of stochastic computing for the embedded system.

Acknowledgment

This study was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program. No. 20006566, ‘Development of embedded artificial intelligence module and system based on neuromorphic’

Reference

[1] J. K. Kim, J. Hwan Oh, J. H. Yang and S. Eun Lee, "2D Line Draw Hardware Accelerator for Tiny Embedded Processor in Consumer Electronics," *IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1-2, 2019.

[2] Armin Alaghi and John P. Hayes, "Survey of Stochastic Computing," *ACM Transactions on Embedded Computing Systems (TECS)*, Volume 12, Issue 2s, pp. 529-551, May. 2013.

[3] P. Li and D. J. Lilja, "Using stochastic computing to implement digital image processing algorithms," *IEEE 29th International Conference on Computer Design (ICCD)*, pp. 154-161, 2011.

[4] K. J. Ahmed, B. Yuan and M. J. Lee, "High-Accuracy Stochastic Computing-Based FIR Filter Design," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1140-1144, 2018.

[5] J. Yu, K. Kim, J. Lee and K. Choi, "Accurate and Efficient Stochastic Computing Hardware for Convolutional Neural Networks," *IEEE International Conference on Computer Design (ICCD)*, pp. 105-112, 2017.

[6] Y. Xie, C. Deng, S. Liao and B. Yuan, "Area-efficient K-Nearest Neighbor Design using Stochastic Computing," *52nd Asilomar Conference on Signals, Systems, and Computers*, pp. 782-786, 2018.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 25,2020 at 20:29:13 UTC from IEEE Xplore. Restrictions apply.