TOKYO UNIVERSITY OF SCIENCE
1-3 KAGURAZAKA, SHINJUKU-KU,
TOKYO 162-8601, JAPAN
Phone: +81-3-5228-8107

May 14, 2018

For promotional use

Reconfigurable Artificial Intelligence (AI) Semiconductor for Advanced use of Information by "Things"
—Sparse ternary neural network AI chip cleverly uses zeros to reduce implementation footprint while increasing recognition accuracy—

Tokyo University of Science

Research Summary

Professor Takayuki Kawahara, of the Department of Electrical Engineering at the Tokyo University of Science, has developed a new method for implementing semiconductors for AI[1] computers (AI chips), which achieves high recognition accuracy using less chip area. The method is simple and effective, achieving high recognition accuracy with neural networks that use three values: 1, 0, and -1. It reduces the number of devices and wiring when implemented on FPGA[2] semiconductor chips by omitting the wiring for 0 values. As a result, recognition accuracy is almost the same as in the ideal case using real-valued weightings, and less area is required than for binary methods, which were thought to use minimal area. This technology is promising for implementing semiconductor computer devices incorporating AI that are suitable for the IoT[3] ("edge") devices with strict power consumption and size constraints, and should contribute to ongoing development of the IoT society and the rapidly increasing amounts of data.

[Research Background]

Technology able to extract meaningful digital information (data) from the environment and societal infrastructure in the real (physical) world is important for sustainable development of the IoT[3] society. As performance of sensor elements improves, the many "things" incorporating these sensors will connect directly to data networks, creating a flood of source data from the real world and placing a heavy load on communication channels and the cloud. Efficient use of electrical power will also decline. To resolve these issues, it will be necessary to process this rapidly increasing amount of data from things at its source, extracting only the useful information. Such processing is of a type that generally requires very large amounts of processing with conventional von Neumann architectures. As such, there is much promise in new data processing equipment based on AI[1]. This field and the field of embedded microcomputers, in which Japan excels, have developed greatly in connection with AI.

Development of neural network (NN) technology has been central to implementation of such AI processing. Compared with conventional computing, NN training requires optimization of parameters such as internal connection weightings and biases. Once these optimal values have been determined,

which is the result of training, recognition tasks can be completed rapidly and accurately by inputting data into the NN. This is called inference, and represents the bulk of processing to utilize data produced by "things". The amount of processing required is an issue with this. Ideally, real values would be used for these parameters, requiring accuracy of about 8 bits in implementation, and many product-sum operations are required to compute NN outputs. This increases circuit size and power consumption. "Things" (at the edges) are much more constrained than servers and data centers in terms of power consumption. As such, research has been conducted on reducing the scale of circuits and memory by restricting such parameters to two values (binary). However, this has not necessarily produced the recognition accuracy required.

Furthermore, there are many types of real world data and many ways it can be processed and used. NN training can handle a wide range of information but to perform inference locally, within "things", NNs must be reconfigurable according to training results. Reconfigurable semiconductors such as FPGAs[2], which are very flexible and can be reconfigured freely, are well-suited to such implementations and can perform inferences with low power and high performance.

Development of AI edge computer technology (AI chip technology) focused on inference is essential for performing this sort of advanced sensor data recognition accurately and with low power consumption within "things" (at the edges), where there are strict constraints on power consumption. Furthermore, such processing is needed to handle a wide range of applications with the next generation of the IoT.

[Overview of Research Results]

Considering the above, we have developed an original sparse ternary neural network AI chip technology using FPGAs. Specific details are described below.

1) A three-valued (ternary) NN expressing weightings with the three values 1, 0, and -1 was used. It was implemented using FPGAs in which weighting values of 0 are implemented by omitting the connections. Conventional binary NNs used two values for weightings, inputs and outputs and were configured using XNOR circuits. However, many connections were still needed between neurons, as shown in Figure 1(a), and problem-dependent schemes had to be used to improve recognition rates. By expressing weightings using three values as shown in Figure 1(b), recognition rates are improved, and weightings of 0 are realized by not implementing the connection. This reduced the size of the implementation. While effectively using three values for weightings, inputs and outputs only use two values.
2) As shown in Figure 2(a), evaluation using the MNIST[4] data set achieved high recognition rates of 93 to 96%, differing in accuracy from ideal results using real-valued weightings by only 0.3 to 0.6%. Also, since weightings of 0 were not implemented in the trial FPGA implementation, as shown in Figure 2(b), there were 46% fewer connections between intermediate and output layers in the NN compared to the binary case, and a 30% reduction in area occupied by memory. Thus, we achieved accuracy comparable to that of the ideal case,

using real-valued weightings, and can expect implementations using less area than with binary weightings, which were thought to minimize area required but have lower accuracy.

Note that we also developed methods for normalizing weightings, setting thresholds and training coefficients, and selecting suitable activation functions for the training needed to implement the ternary neural networks.

These research results were presented at the IEEE International Symposium on Next-Generation Electronics (ISNE 2018), held on May 7-9, 2018 in Taipei.

[Future Prospects]

- Demonstrate size reduction and accuracy on a real device by building a demo system.
- Demonstrate using a real application

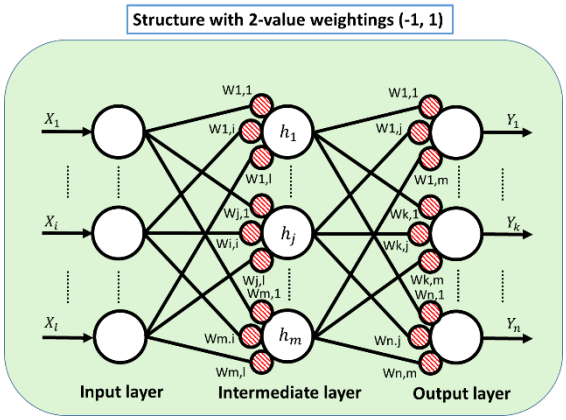

Figure 1(a) Conventional method

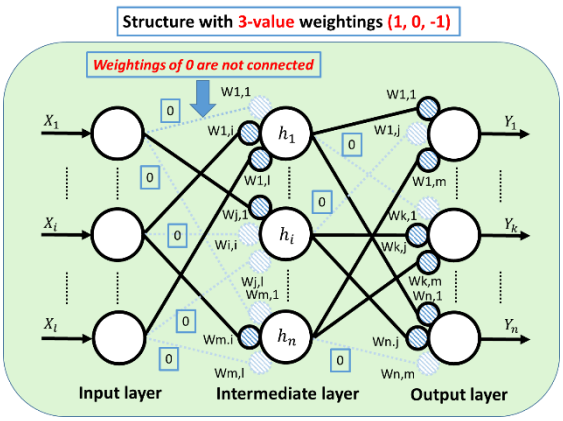

Figure 1(b) Proposed method

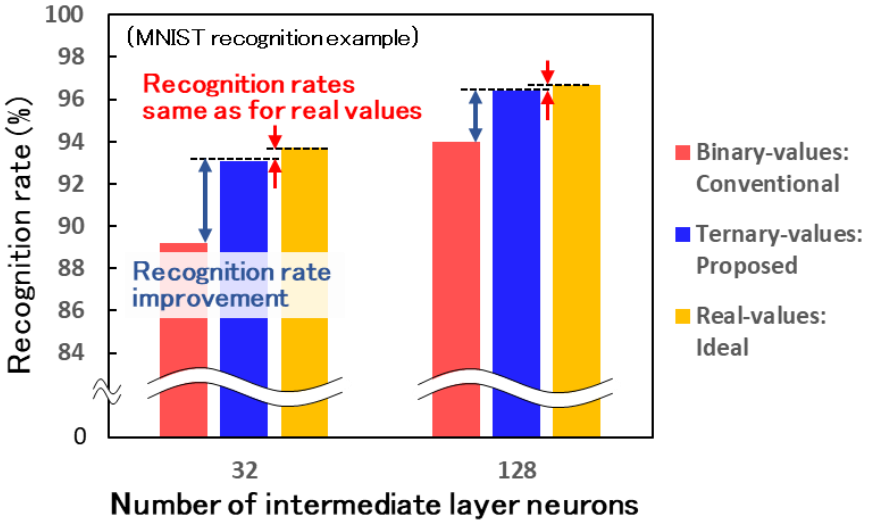

Figure 2(a) Recognition rate improvements

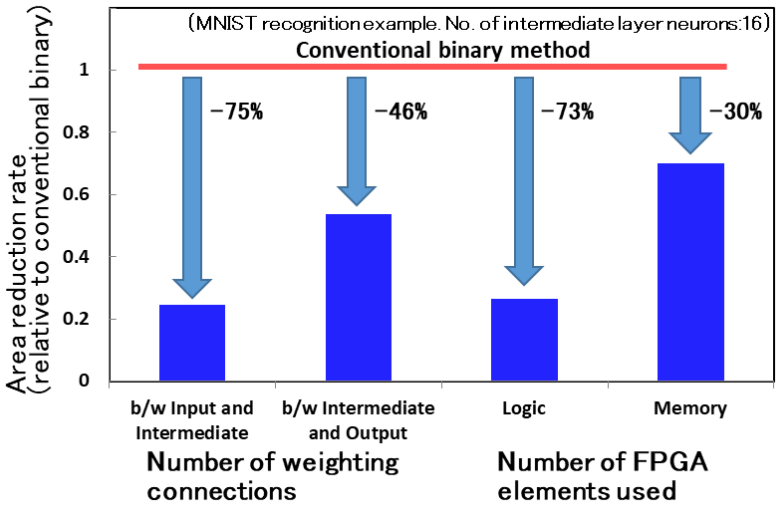

Figure 2(b) Chip-area reduction effect

Terminology

1 AI: Artificial Intelligence.

2 FPGA: Field Programmable Gate Array. A type of LSI containing logic circuits that can be combined and rearranged after the LSI has been fabricated to implement the desired functionality.

3 IoT: Internet of Things. The addition of Internet functionality to various objects (“things”). Objects that previously could not connect directly to the Internet are given the ability to exchange information over the Internet.

4 MNIST: A data set of images of hand-written digits from 0 to 9. There are 6000 handwritten images for each of the digits. They are used to evaluate neural network learning and inference methods.

-- Inquiries --
University Research Administration Center
TOKYO UNIVERSITY OF SCIENCE
1-3, Kagurazaka, Shinjuku-ku, Tokyo, 162-8601, Japan
E-MAIL: ura@admin.tus.ac.jp