2023 IEEE 12th Global Conference on Consumer Electronics (GCCE) | 979-8-3503-4018-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/GCCE59613.2023.10315393

An Edge AI Accelerator Design Based on LRCN Model for Real-time EEG-based Emotion Detection System on the RISC-V FPGA Platform

Jia-Yu Li, Yi-Kai Chen, Wai-Chi Fang
Institute of Electronics
National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan
Corresponding Author: Professor Wai-Chi Fang (wfang@nycu.edu.tw)

***Abstract*—With the development of neural networks and big data, research on emotion recognition has gradually increased. In many ways of emotion recognition, we proposed using physiological signals such as Electroencephalography(EEG) to achieve high-precision emotion recognition. In this paper, the model we built was based on the concept of Long-term Recurrent Convolutional Networks(LRCN), which is used for emotion recognition of EEG signals. In addition, we achieved better performance in terms of area and speed by introducing and optimizing an AI architecture with high data reuse. Our structure adopt a strategy of high data reusability and can be reprocessed to various size of convolution layer and fully connected layer. The optimized architecture can achieve an peak performance of 128.2 GMACS/W, and uses only 196 KGates (NAND2) with average power consumption 0.039 mW in TSMC 90-nm.**

***Index Terms*—Affective Computing, Electroencephalogram, Accelerator, LRCN, Deep Learning**

I. Introduction

Emotion computation has become increasingly popular in the realm of Human-Computer Interaction (HCI) design. Physiological signals are widely regarded as the most objective and dependable means of identifying human emotions. Numerous studies [1] have demonstrated that electroencephalogram (EEG) serves as a data source with superior classification accuracy and rapid responsiveness to emotional stimuli, surpassing other physiological signals. As a result, we have previously devised a portable emotion computation system [2], [3] in our research endeavors. However, wearable medical sensors and IoT devices monitor a majority of biomedical signals, but they have limitations in terms of size and power consumption [4]. Hence, optimizing deep learning through appropriate hardware design, data flow, and compression is indispensable.

In order to reduce power consumption, We adapted and optimized our previous hardware by modifying Eyeriss [5]. They proposed row stationary which optimize data reuse maximally to meet convolution algorithm and reduce the power of data transfer from external memory. According to Eyeriss, although it only focused on the algorithm of convolution, we thought that fully connected layer also has data reusability, which means it can use this hardware architecture to perform operations to a certain extent. Therefore, we propose a reconfigurable neural network accelerator architecture which can fit various sizes of CNN model and ANN model.

II. System Architecture

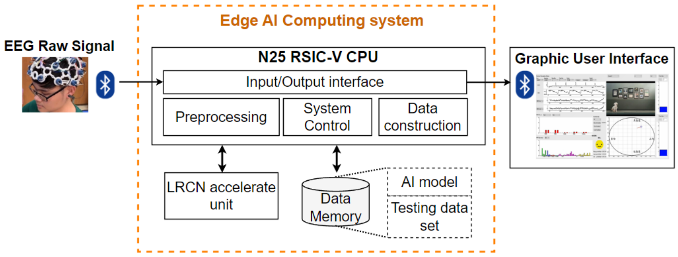

Fig. 1. Real-time EEG-based Emotion Recognition System.

The architecture of the emotion computing platform we propose, depicted in Figure 1, consists of three main components: an EEG wearable front-end sensor, an AI FPGA platform, and a laptop computer with a graphical user interface. The EEG front-end sensor captures raw data, which is transmitted to the FPGA platform through Bluetooth for signal preprocessing and feature extraction. The processed data is then forwarded to the LRCN accelerator, where the emotion classification takes place. The controller within the AI platform receives the classification outcome from the accelerator and communicates it to the graphical user interface on the laptop computer via Bluetooth.

III. Emotion Recognition Method

A. RISC-V based Data Processing

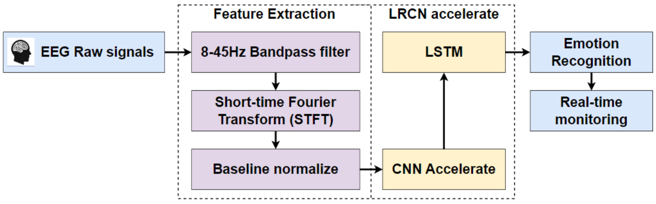

Fig. 2. RISC-V based Data Handling Processing.

Figure 2 shows the EEG signal processing flow. The RISC-V processor performs preprocessing and emotion feature extraction on the received signals. For signal preprocessing, we selected an 8-channel EEG signal due to computational limitations of the RISC-V processor. According to the research conducted by Yang et al. [6], the following channels, namely Fp1, Fp2, F3, F4, F7, F8, T3, and T4, were suggested as suitable choices for EEG-based emotion recognition algorithms. After computing the emotion features, the RISC-V processor controls their transmission to the emotion recognition unit for LRCN algorithm computation. We propose a CNN acceleration architecture to speed up the LRCN architecture, utilizing high data reuse to reduce memory accesses. Finally, the LRCN produces emotional outputs and undergoes validation.

B. *Edge AI Accelerator based on LRCN*

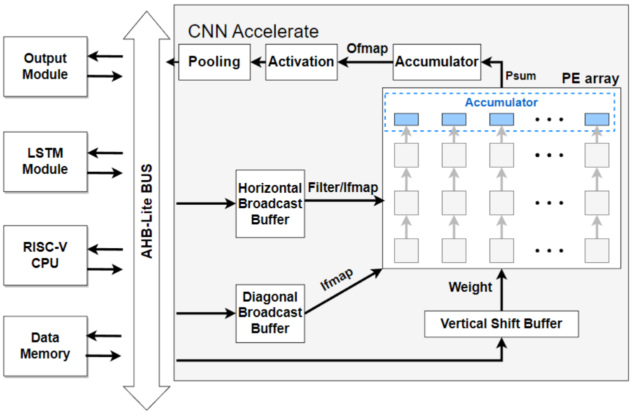

Fig. 3. The architecture of LRCN accelerate with processing engine array.

The LRCN model which is proposed in our previous work [2] consists of 3 convolution layers, 3 average pooling layers, a LSTM layer, a fully connected layer, and a softmax layer. Regarding this architecture, we have optimized the convolution and fully connected operations. Figure 3 illustrates the optimized architecture. We propose two data flows, namely the broadcast data flow and the shift data flow. The broadcast data flow consists of two paths: horizontal and diagonal, which is consistent with [5]. We use the diagonal broadcast input and the horizontal broadcast input to do the convolution operation. For the fully connect, we use horizontal broadcast for the input feature map and vertical shift register for the weights. Through this architecture, we have integrated two operations into the same hardware, achieving optimization results.

IV. Experiment Result

To enable a meaningful comparison with prior findings, we deploy the AI Accelerator using the Synopsys Design Compiler utilizing the TSMC 90-nm technology. The estimation of chip-only power consumption is performed by Cadence Innovus after conducting post-simulation analysis. The architecture uses only 196 KGates (NAND2) to implement the hardware and the peak performance can reach 128.2G operation per second when the core frequency is 170 MHz under TSMC 90-nm.

Because our primary focus was on optimizing the convolution and fully connected operations, we compared these two operations using this simulation environment. The comparison shows in table I. The latency of processing 3 different sizes of convolution layers in our structure is averagely faster than LRCN chip. In power analysis, the reason for our higher energy consumption is to accommodate sufficient SRAM usage for different applications, with energy consumption outside of SRAM accounting for only 30% of the total. For fully connected operations, we sacrifice a slight computational speed to achieve lower energy consumption and share a significant portion of hardware with convolution operations. This architecture enables other biomedical-related AI applications and demonstrates good energy efficiency for one-dimensional inputs such as signal recognition.

TABLE I
Comparison of our structure with LRCN chip

Layer	LRCN [2]		ours	
	Latency(cycle	Power(mW)	Latency(cycle	Power(mW)
CONV1	308	0.0161	391	0.0228
CONV2	278	0.008	199	0.0205
CONV3	230	0.0105	167	0.0203
FC	4	0.0031	19	0.0029

V. Conclusion

In our previous research, we optimized and proposed a new architecture that is lightweight and integrates convolutional and matrix-vector multiplication operations. The architecture offers advantages stemming from the energy efficiency of the paper [5]. We've further enhanced and integrated it for optimal use in wearable biomedical devices. As validated by the experimental results, the proposed AI processor not only has the advantage of high computing performance for targeted AI-inspired biomedical applications but also maintains high energy efficiency.

References

[1] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang, "Accurate eeg-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 44 317–44 328, 2019.

[2] C.-J. Yang, W.-C. Li, M.-T. Wan, and W.-C. Fang, "Real-time eeg-based affective computing using on-chip learning long-term recurrent convolutional network," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[3] C.-J. Yang, N. Fahier, C.-Y. He, W.-C. Li, and W.-C. Fang, "An ai-edge platform with multimodal wearable physiological signals monitoring sensors for affective computing applications," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[4] K.-Y. Wang, Y.-D. Huang, Y.-L. Ho, and W.-C. Fang, "A customized convolutional neural network design using improved softmax layer for real-time human emotion recognition," in *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2019, pp. 102–106.

[5] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.

[6] W.-C. Li, C.-J. Yang, B.-T. Liu, and W.-C. Fang, "A real-time affective computing platform integrated with ai system-on-chip design and multimodal signal processing system," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 522–526.