2023 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-0026-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/BioCAS58349.2023.10389052

An Edge AI Accelerator of LRCN Model with RISC-V Platform for EEG-based Emotion Real-time Detection System

Yi-Kai Chen, Jia-Yu Li, 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 Electroencephalography (EEG) signals to achieve high-precision emotion recognition. In this paper, the model we built was based on the concept of Long-term Recurrent Convolution Networks (LRCN), which is used for emotion recognition of EEG signals. In order to realize this real-time wearable system, we employ RISC-V for signal preprocessing and establish the entire system by communicating with our AI accelerator through a communication protocol. In addition, to accelerate and integrate this AI architecture into the RISC-V platform, we optimized the area and computing efficiency of the AI architecture. This optimization improves the data reuse of convolution and fully connected operations and enables acceptance of inputs of different sizes, maximizing hardware reusability. Finally, the AI acceleration chip within the system was implemented on the Kintex-7 platform, achieving an accuracy of 88.6% (two-class classification) and 69.31% (three-class classification) on the SEED dataset and the optimized AI architecture exhibits a power efficiency of 9.26 GOPS/W.**

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

I. Introduction

Emotion computation has become increasingly popular in the realm of Human-Computer Interaction (HCI) design. The method of emotion recognition can be analyzed from external facial expressions [1], gestures, speech tones, and also from human physiological signals such as EEG, EMG, PPG, ECG, etc. However, some fields have higher requirements for the accuracy of the emotion recognition system, such as medical applications. In this case, the accuracy of emotion recognition using facial expressions, gestures and speech tones is not enough because it varies from cultures and nations [2]. In this situation, physiological signals are widely regarded as the most objective and dependable means of identifying human emotions. Many research studies [3] have shown that electroencephalogram (EEG) data is a valuable source for accurately classifying emotions and quickly responding to emotional stimuli, superior to other physiological signals. In order to obtain a more objective and higher accurate emotion recognition result, we used physiological signals for recognition.

As a result, we previously designed a portable affective computing system in our research work [4], [5]. However, wearable medical sensors and IoT devices monitor a majority of biomedical signals, but they have limitations in terms of size and power consumption [6]. 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 [7]. Eyeriss is a novel design which process configurable convolution neural network for energy efficiency by maximally reusing data. 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 hardware-optimized AI accelerator to enhance the energy efficiency of our real-time EEG emotion recognition system.

II. System Architecture

The architecture of our proposed emotion computing platform, depicted in Figure 1, comprises three main components: an EEG wearable front-end sensor, an AI acceleration chip that can communicate with the RISC-V platform, and a laptop computer with a graphical user interface(GUI).

In our emotion recognition system, the EEG front-end sensor captures raw data related to brain activity first. The raw data is wirelessly transmitted via Bluetooth to the RISC-V platform for signal preprocessing and feature extraction. The

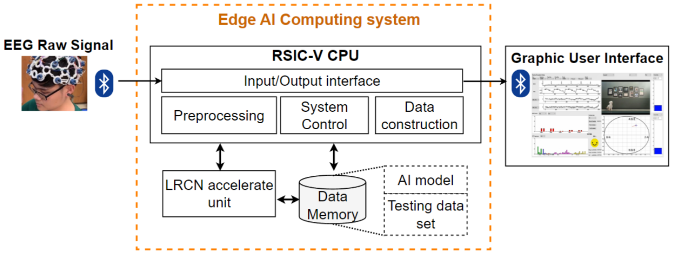

Fig. 1. Real-time EEG-based Emotion Recognition System.

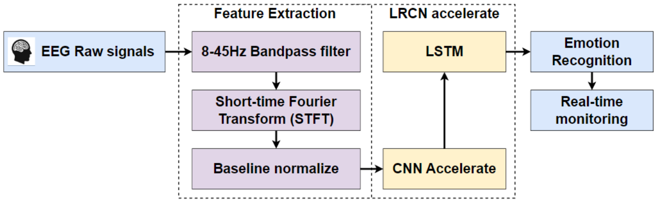

Fig. 2. Data processing of EEG-based emotion recognition system.

processed data is then forwarded to the LRCN accelerator, responsible for analyzing and classifying the user's emotions. The controller within the RISC-V platform receives the classification result from the AI accelerator and transmit to the GUI on the laptop computer via Bluetooth.

III. System Data Processing

Figure 2 illustrates the data processing of our system. In our system, there are two main data computation components: feature extraction and AI algorithms based on the LRCN model. These two types of data computations are respectively implemented by the RISC-V CPU and the AI accelerator.

A. Feature Extraction of Raw Data

The RISC-V processor are mainly performs pre-processing stage and the extraction of emotion-related features from the received signals. We reduce the number of utilized channels to 8 in order to decrease the computational load on the RISC-V processor. According to previous research conducted by Yang et al. [8], the selected channels for EEG-based emotion recognition algorithms are Fp1, Fp2, F3, F4, F7, F8, T3, and T4, and each EEG channel having a resolution of 24 bits. The sampling frequency for each channel is set at 256 samples per second.

For signal feature extraction, based on our previous research experience [8], [9], we have observed that the Alpha (8-13 Hz), Beta (14-30 Hz), and Gamma (31-45 Hz) frequency bands of the EEG signals exhibit significant features for emotion recognition. Therefore, we initially apply a bandpass filter to the signals in the range of 8-45 Hz. Subsequently, we perform short-time Fourier transform (STFT) on this frequency band using a window size of 1 second with no overlap. Finally, we employ baseline normalization [10] to eliminate inter-individual variations and further enhance the accuracy of emotion recognition.

B. Algorithm of LRCN Model

The LRCN model which is proposed in our previous work [4] consists of 3 convolution layers, 3 average pooling layers, a LSTM layer, a fully connected layer, and a softmax layer. In the LRCN model, we employ a many-to-one architecture for recognition. The current output is determined based on the current input and the previously recorded nine seconds of input. Due to the complexity of this architecture, we have specifically optimized the computationally intensive convolution and fully connected layers in hardware to enhance the overall efficiency of the system.

IV. Hardware Implementation

A. The Architecture of Edge AI Accelerator

The proposed system of emotion recognition was developed as shown in figure 3. Data transmission between the AI acceleration unit and RISC-V in the hardware implementation is facilitated using the APB communication protocol. However, only model parameters, quantized inputs, and recognition results need to be transmitted. The intermediate results calculated by each layer of the AI model will be stored in the shared data SRAM. The overall architecture depicted in the diagram shows that the AI accelerator is divided into three modules: the optimized CNN module, the shared data SRAM, and the LSTM module. The CNN module contain the computations required for the convolution layer and fully-connected layer. The shared DATA SRAM consists of five double-word SRAMs to provide the bandwidth required for computational modules to process 10 kernels or units. The RISC-V controls the modules and their access permission to SRAM through APB.

B. The Optimized CNN module

We optimize the computations needed for the convolution layer and fully connected layer in the direction of hardware sharing and data reuse, aiming for reduced area and improved energy efficiency. In our optimized CNN accelerator, 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 [7]. In Figure 4, these two data flows are labeled as Conv and FC. During convolutional operations, the filter and input feature map are transmitted to multiple Processing Engines (PE) using horizontal broadcasting and diagonal broadcasting, respectively. During fully connected operations, weights are similarly transmitted to multiple PE using horizontal broadcasting, while the input feature map is pushed to a single PE engine in one data at a time using vertical shifting.

The main operational of the CNN accelerator consists of multiple PE. The design of the PE architecture is illustrated in Figure 5. Every three PEs form a column in the unit PE array, which operates accumulation in an upward direction. Horizontal and diagonal inputs are broadcasted through a broadcast buffer, and vertical inputs are sequentially shifted. Additionally, we use multiplexers to select one of two data flow, vertical shifting and diagonal broadcasting. The top-level component of the PE is the accumulator, which can accumulate values across the entire column. Furthermore, we employ 16-bit fixed-point arithmetic in Q8.8 format since it provides sufficient precision for inference and is more energy-efficient than floating-point arithmetic.

V. Experiment Result

The physical diagram of the real-time system in this study is shown in Figure 6. As mentioned in Section II, the equipment used includes an 8-channel EEG front-end sensor, an AI acceleration chip with RISC-V CPU platform, and a laptop computer for real-time GUI display. The hardware

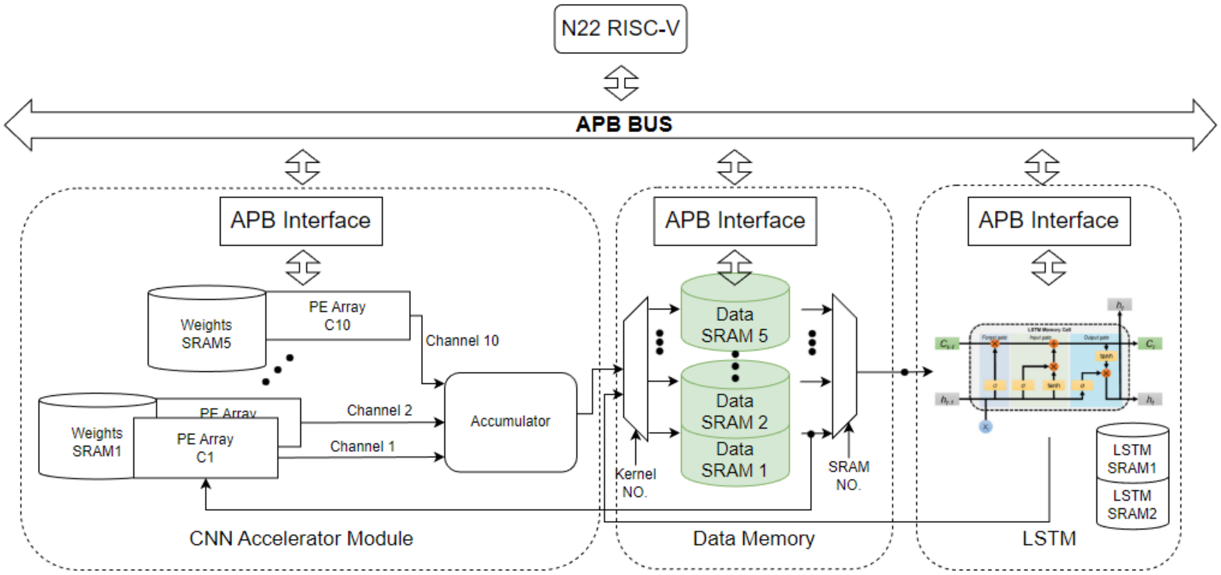

Fig. 3. The architecture of AI accelerator based on LRCN model.

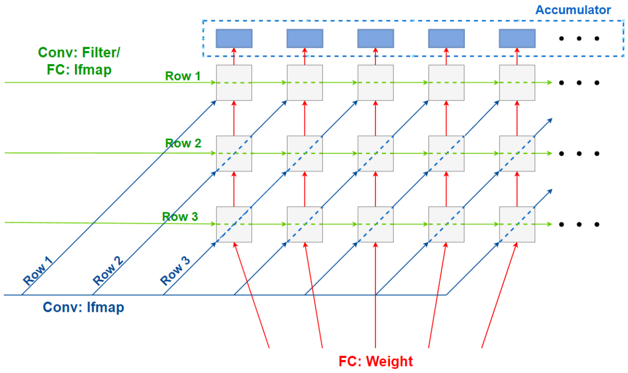

Fig. 4. PE array with two type of data flow.

environment is built on the Xilinx Kintex-7 xc7k160t FPGA, and hardware synthesis is performed using Vivado 2021.2. Bluetooth devices are used for data transmission between the devices.

A. Comparison of Accuracy

Figure 7 presents the emotion recognition results obtained using the SEED dataset with the AI architecture based on LRCN. The dataset consists of 15 subjects, and Figure 7(a) shows the two-class classification result, which achieved an accuracy of 88.6%. Figure 7(b) shows the results for the three-class classification (positive-negative-neutral), with an accuracy of 63.91%.

A comparison with other state-of-art research is shown in Table I. Although our system does not demonstrate outstanding performance in terms of accuracy compared to the studies referenced [11]–[13], it is worth noting that these studies did not implement their models as hardware systems. Therefore, the model architectures used in these previous studies may impose greater hardware burden. The architecture proposed in reference [14] achieved a complete hardware implementation, but its accuracy rate is slightly lower than our research. The results indicate that, in the context of system implementation, our study is better than many of the compared research studies in terms of accuracy.

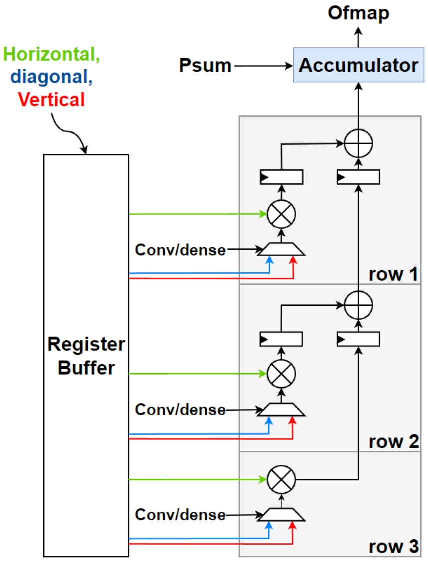

Fig. 5. One column of PE engine in PE array

B. Comparison of Hardware Implementation

A comparison with other state-of-art research implemented on hardware is shown in Table II. The data displayed in the table represents the results after placement and routing of the AI accelerator unit only. Among these studies, only reference

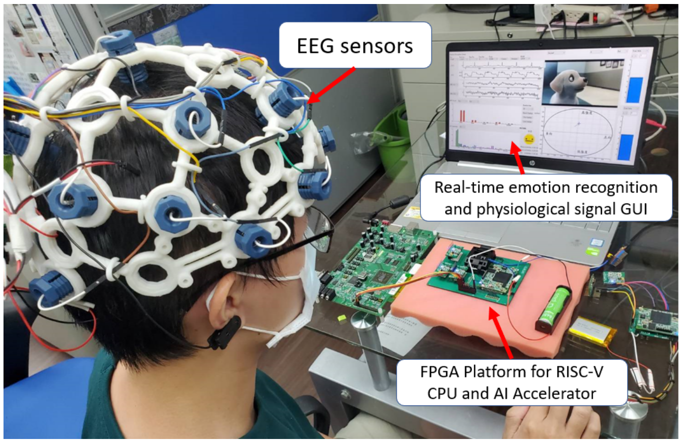
EEG sensors
Real-time emotion recognition and physiological signal GUI
FPGA Platform for RISC-V CPU and AI Accelerator

Fig. 6. Real-time emotion recognition system based on LRCN on RISC-V FPGA platform.

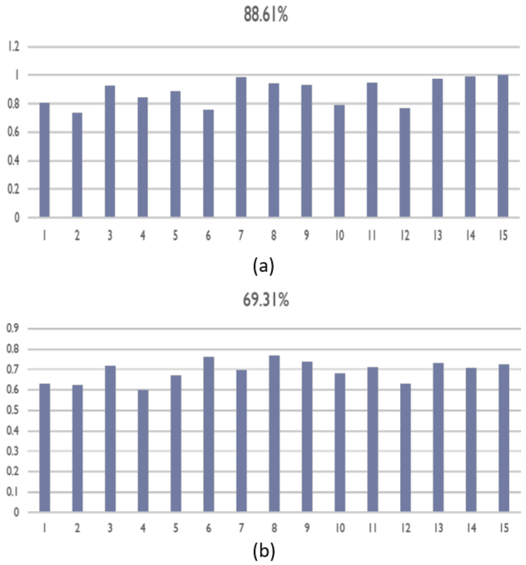

Fig. 7. The emotion classification results of 15 subjects on SEED dataset. (a)2-class. (b)3-class.

[14] shares the same research direction as ours, focusing on the implementation of hardware systems for emotion recognition. In comparison to [14], our approach includes LSTM computations, which were not specifically optimized, resulting in an impact on our final energy efficiency results. However, the LRCN architecture demonstrates outstanding accuracy in recognition, which is the main reason for its superiority in accuracy compared to the research [14]. Furthermore, we achieve a less latency of only 0.014 ms for each input. This is attributed to the fact that we reduce and select meaningful EEG channels to reduce the amount of data and achieve less latency results. In general, our hardware architecture still exhibits remarkable energy efficiency among other advanced research studies. Finally, incorporating RISC-V into the entire hardware system for placement and routing resulted in the utilization of 29,438 LUTs, 16,313 FFs, and 122 BRAMs.

TABLE I
Accuracy Comparison With orther works

	[11]	[12]	[13]	[14]	ours
Dataset	SEED	SEED	DEAP	DEAP	SEED
EEG channels	62	9	32	14	8
Classifier	ST-SBSSVM	ResNet	DBCN	BioCNN	LRCN
Class	2	2/3	2	2	2/3
Accuracy	89.0%	86.6% -2 78.3% -3	83.98%	77.6%	88.6% -2 69.3% -3

TABLE II
FPGA hardware comparison with orther works

	[14]	[15]	[16]	ours
Clock	100 M	86.73 M	50 M	50 M
Platform	Atlys	Virtex-5	Artix-7	Kintex-7
LUTs	26229	11963	65731	12364
FFs	15180	-	35199	7544
BRAMs	10	40	68	20
Power	0.15 W	1.59 W	0.11 W	0.18 W
Energy efficency	11.0 GOPs/W	-	8.19 GOPs/W	9.26 GOPs/W
Latency	0.93 ms	-	-	0.014 ms

VI. Conclusion

In this paper, we optimized previous research and proposed a new lightweight architecture which is used to integrate convolution and matrix-vector multiplication operations. This architecture primarily optimizes data reuse and hardware sharing to achieve high energy efficiency, making it well-suited for application in wearable devices for emotion recognition systems. This gives our AI accelerator an accuracy of 88.6% under Kintex-7 platform with an energy efficiency of 9.26 GOPs/W. Furthermore, the entire system combines the RISC-V CPU to operate feature extraction computations and controls the AI accelerator. We believe that such a hardware system construction can not only achieve high energy efficiency in research related to emotion recognition, but also achieve outstanding results in AI-related biomedical applications.

References

[1] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/2/401

[2] J. Tao and T. Tan, "Affective computing: A review," in *Affective Computing and Intelligent Interaction*, J. Tao, T. Tan, and R. W. Picard, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 981–995.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] W.-C. Fang, K.-Y. Wang, N. Fahier, Y.-L. Ho, and Y.-D. Huang, "Development and validation of an eeg-based real-time emotion recognition system using edge ai computing platform with convolutional neural network system-on-chip design," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 4, pp. 645–657, 2019.

[10] M. X. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press, 2014.

[11] F. Yang, X. Zhao, W. Jiang, P. Gao, and G. Liu, "Multi-method fusion of cross-subject emotion recognition based on high-dimensional eeg features," *Frontiers in Computational Neuroscience*, vol. 13, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2019.00053

[12] Y. Cimtay and E. Ekmekcioglu, "Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset eeg emotion recognition," *Sensors*, vol. 20, no. 7, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/7/2034

[13] Q. Yao, H. Gu, S. Wang, and X. Li, "A feature-fused convolutional neural network for emotion recognition from multichannel eeg signals," *IEEE Sensors Journal*, vol. 22, no. 12, pp. 11 954–11 964, 2022.

[14] H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel, "Biocnn: A hardware inference engine for eeg-based emotion detection," *IEEE Access*, vol. 8, pp. 140 896–140 914, 2020.

[15] M. Sahani, S. K. Rout, and P. K. Dash, "Epileptic seizure recognition using reduced deep convolutional stack autoencoder and improved kernel rvfln from eeg signals," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 3, pp. 595–605, 2021.

[16] C. Zhang, Z. Tang, T. Guo, J. Lei, J. Xiao, A. Wang, S. Bai, and M. Zhang, "Salenet: A low-power end-to-end cnn accelerator for sustained attention level evaluation using eeg," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022, pp. 2304–2308.