RAISE: A Resistive Accelerator for Subject-Independent EEG Signal Classification

Fan Chen, Linghao Song, Hai "Helen" Li, Yiran Chen
Department of Electrical and Computer Engineering, Duke University, Durham NC, U.S.A.
Email: {fan.chen, linghao.song, hai.li, yiran.chen}@duke.edu

***Abstract*—State-of-the-art deep neural networks (DNNs) for electroencephalography (EEG) signals classification focus on subject-related tasks, in which the test data and the training data needs to be collected from the same subject. In addition, due to limited computing resources and strict power budgets at edges, it is very challenging to deploy the inference of such DNN models on biological devices. In this work, we present an algorithm/hardware co-designed low-power accelerator for subject-independent EEG signal classification. We propose a compact neural network that is capable to identify the common and stable structure among subjects. Based on it, we realize a robust subject-independent EEG signal classification model that can be extended to multiple BCI tasks with minimal overhead. Based on this model, we present RAISE, a low-power processing-in-memory inference accelerator by leveraging the emerging resistive memory. We compare the proposed model and hardware accelerator to prior arts across various BCI paradigms. We show that our model achieves the best subject-independent classification accuracy, while RAISE achieves 2.8× power reduction and 2.5× improvement in performance per watt compared to the state-of-the-art resistive inference accelerator.**

***Index Terms*—BCI, EEG, DNN, resistive memory**

I. Introduction

A brain-computer interface (BCI) identifies intents from human brain activity patterns and converts them into commands for an interactive application, thereby enabling direct communication between human brain and external devices. Recent research on BCI system has demonstrated the potential to revolutionize medical applications [1], [2], as well as non-medical areas [3], [4]. Figure 1 illustrated the basics of BCI systems. Neural signals are first collected through a wearable headset with multiple electrodes (*i.e.,* channels). The collected signal presents different waveform shapes can be generally categorized into two paradigms: event-related (e.g., event-related potential (ERP)) and oscillatory (*e.g.*, sensory motor rhythm (SMR)). Movement-related cortical potential (MRCP) can be regarded as a mixture of the two because it contains both components. The BCI processing normally consists of five stages, among which the the accuracy of BCIs mainly depends on feature extraction and classification. Previous work [5]–[7] have mainly been applied to within-subject tasks, and cannot be generalized to cross-subject tasks. Moreover, as the number of channels in BCI systems is increasing rapidly, the corresponding computing throughput and power consumption is hard to satisfied under the limited budget in biological devices.

A. Deep Learning Models for EEG Signal Classification

Neural signals exhibit distinct characteristics. ERP BCIs detect high-amplitude (∼10 uV) and low-frequency (<40 Hz) neural responses to a time-locked stimulus. SMR BCIs capture the signal power within a specific frequency band. MRCP BCIs capture both amplitude perturbations and oscillatory components. Conventional EEG signal processing flow relies on band-pass filters, handcrafted feature extractors and classification methods, which suffers from low accuracy, high processing energy, and low generalization capability when subject changes.

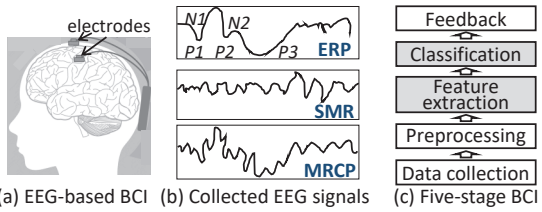

Fig. 1. EEG-based BCI basics.

Recent efforts explored the application of deep learning to EEG [5]–[7], demonstrating significantly enhanced accuracy. The main stream deep learning models used in EEG classification can be categorized into two types: (i) full convolutional neural networks (CNN) [5], [6]; (ii) integrated convolutional neural networks and long short-term memory (LSTM) models [7]. As illustrated in Figure 2 (a), LSTM-based models learn the frequency and spatial filters using CNNs, while learn temporal representations using LSTM layers. The input raw data collected from electrodes are preprocessed into 2D matrix before feeding into the model. For instance, [7] converts every 64 sampling data to a 10×11 topology-preserving matrix through zero insertion, Hence, >40% of the computation is multiplication or addition with zero operands, resulting significant resource underutilization. The recent proposed EEGNet [5] eliminates such data reorganization and processes the raw data samples directly. As shown in Figure 2 (b). EEGNet is constructed mainly from separable CONV layers consisting of depthwise CONV (dCONV) and pointwise CONV (pCONV).

B. Transfer Learning

Transfer learning [8] leverages a previous trained neural network and adapts it to a new task by reusing the common features at initial layers while allow the model to learn specific adaptation layers for the target task. To improve model efficiency by supporting multiple BCI tasks, We consider sharing common structures in different EEG tasks (if any).

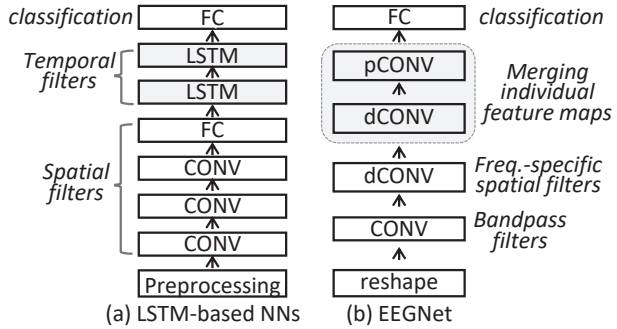

Fig. 2. Deep learning models for EEG classification.

TABLE I
NETWORK ARCHITECTURE FOR A 64-CHANNEL BCI SYSTEM WITH 128HZ SAMPLING RATE.
(D = depth of the channel multiplier in depthwise convolutional layers, N = number of classes)

Block	Layer	Filter					Output			Activation	Options
		#	Channel	Height	Width	Stride	Channel	Height	Width		
Input							1	64	128		
1	CONV2D	4	1	1	64	1	20	64	128	ReLu	SAME padding
		4	1	1	16	1					
		4	1	1	8	1					
		4	1	1	4	1					
2	dCONV	32	1	64	1	4	32	1	32	ReLu	VALID padding D=2
3	dCONV	32	1	1	16	4	32	1	8	ReLu	SAME padding D=1
	pCONV	16	32	1	1	1	16	1	8	ReLu	SAME padding
Classifier	dense	N*(16*8)					N	1	1	Softmax	

Our exploration on EEGNet [5] shows that the initial kernels are approximating conventional bandpass filter functions. After the EEG signals have been bandpass-filtered into multiple frequency bands, feature extracting kernels are then applied to learn discriminative features for specific BCI paradigms. Therefore, we present to explores the similar transferability of deep learning models in performing different EEG tasks.

II. A SUBJECT-INDEPENDENT NETWORK ARCHITECTURE

A. Proposed Network Architecture

The proposed network is based on EEGNet [5]. The original model has three main blocks: (i) block 1 learns frequency filters using 2D CONV (CONV2D); (ii) block 2 uses dCONV to learn frequency-specific spatial filters; (iii) block 3 is a combination of dCONV and pCONV. dCONV learns a temporal summary for each feature map individually, while pCONV learns how to optimally mix the feature maps together.

We investigate the neurophysiologically interpretable features learned in each block and observed that EEGNet can provide better than previous cross-subject classification mainly depends on: (i) independently learning representations on frequency, spatial, and temporal allows the model to target one goal at each block. Therefore, the essential common structures can be captured precisely in each corresponding block; (ii) high level features are generated from combinations of robust low-level features, resulting highly adaptive inference capability when subjects change. Based on these observation, we modify the original network to enable subject-independent EEG signal classification. The key idea is to fully decouple the learning of frequency, spatial, and temporal representations. At initial layer, we fully separate frequency bands of the interest. While later layers gradually embed more subject- or paradigm-specific features. A full description of the modified model for a 64-channel BCI system with T time points is shown in Table I. Below summarizes the key modification on model architecture:

- EEGNet fixes the CONV2D filters size to be half the sampling rate, which allows the output feature maps capture frequency bands at 2 Hz and above. We increase the number of CONV2D kernels with various kernel sizes, leading to more fine-grained frequency filters with enhanced network capacity. Specifically, CONV2D kernels with sizes of (1, 64), (1, 16), (1,8), and (1, 4) capture frequency information on >2 Hz, >8 Hz, >16 Hz, and >32 Hz, respectively.
- We replace the liner convolutions in EEGNet with nonlinear activation. Our experimental results show that using nonlinear activation indeed provides no performance gains in within-subject tasks as observed in EEGNet, however, a significant accuracy improvement is achieved (∼6%) in cross-subject tasks. A very reasonable explanation is that adding non-linearity allows the model to create complex mappings between multiple frequency bands, which are essential for identifying the reusable structure in different subjects while also maintaining adaptation to the variations among subjects.
- We replace the deterministic spatial average pooling layers in block 2 and 3 with strided convolution [9], allowing the network to learn its own spatial downsampling.
- We leverage the similar concept of transfer learning [8] to investigate whether the model can be applied in cross-paradigm tasks. The key approach is to re-utilize initial layers while allowing later layers to be paradigm-specific. Specifically, we share the trained parameters in block 1 and block 2 for low-level representation learning, while maintaining three instances of paradigm-specific block 3 respectively for BCIs focusing on ERP, SMR and MRCP dataset.

B. Training Approach

We use the Adam optimizer with same default parameter in [5]. The model is first trained with PhysioNet EEG Motor Movement/Imagery Dataset [10]. We conducted a 400-epoch training with a 0.1 training rate for the first 200 epochs, 0.01 for the following 100 epochs, and 0.001 for the last 100 epochs. The batch size is 16. The pre-trained model is then used as the starting point for different single-paradigm training. For cross-paradigm tasks, we fix the parameters in block 1 and block 2 in the pre-trained model, only perform parameter updating for the last separable convolution layers. We use floating-point data precision for model training, and quantized 8-bit fixed-point data precision for inference. All models are trained on an NVDIA Geforce GTX 1080 GPU in Pytorch [11].

III. CIRCUITS AND IMPLEMENTATION

Overall architecture. the overall architecture of RAISE is illustrated in Figure 3 (a) . It leverages resistive memory (ReRAM) for both storage and in-memory matrix-vector multiplications. Batch normalization unit (BN) and activation unit (Act.) are customized circuitry for batch normalization and nonlinear activation. As shown in Figure 3 (b), each processing unit (PU) consists multiple ReRAM crossbars (XB)

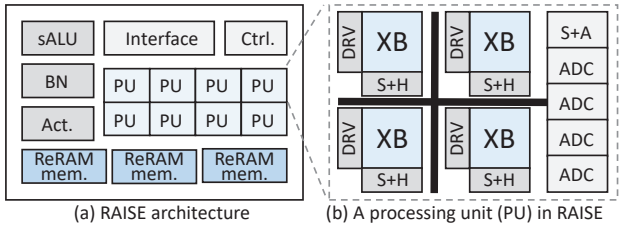

Fig. 3. RAISE architecture hierarchy.

equipped with wordline drivers (DRV) and sample-and-hold (S+H) connected through on-chip mesh. The aggregated output from ReRAM crossbars are connected with analog-to-digital converter (ADC) and shifted-and-add (S+A) circuits. Convolutions are transformed to matrix multiplication through the Toeplitz matrix, then mapped onto PUs, and processed in a pipelined fashion. The processing flow and functionality of these peripheral circuits are standard in current resistive accelerators, we omit the details of these techniques and refer interested readers to related works [12]–[14].

Reduced ADC overhead. Previous DNN inference accelerators targeting large-scale networks require high-speed ADC for real-time processing. ISAAC [12] adopts a single 1.28 giga-samples-per-second (GSps) ADC for each resistive crossbar. However, bio-embedded applications usually employ DNN models with thousands to hundreds of thousands parameters, working at several hundred hertz. In this work, we reduce the ADC performance to accommodate the requirement of BCI tasks. Wwe employ a 50nW 5kHz bandwidth ADC design [15] for the targeted 64-channel BCI systems with 128Hz sample rate. Unlike ISAAC that shares an ADC between 128 bitlines, RAISE shares each ADC among 4 bitlines, corresponding to a 0.8 ms cycle time, which meets the system requirements.

Simplified on-chip communication. Current resistive accelerators are implemented in a tile-based structure. Each tile contains multiple processing engines. Multiple tiles communicate via on-chip connection. Overall, a significant portion of the chip area is occupied by connections. Since the EEGNet-level small network can fit within several PUs, we implement RAISE in a two-level hierarchy instead of a three-level architecture as in ISAAC. The area occupied by on-chip connections accordingly reduced from ∼30% in ISAAC to less than 1% in RAISE.

Design implementation. We report the power and area overhead of RAISE in Table II. For fair comparison with ISAAC [12], we adopt its ReRAM device parameters, DAC, S+H and S+A circuits. The bus and connections, activation, and batch normalization units are synthesized by Cadence Virtuoso with TSMC 32nm process technology. The parameters of ADC is adopted from [15] and scaled to 32nm.

TABLE II
THE HARDWARE COST OF RAISE.

Name	Param.	Spec.	Power (mW)	Area (mm^2)
ReRAM mem.	size	1024×1024	147	0.0117
	number	8×		
Activation	number	1×	0.26	0.0003
BN	number	1×	0.18	0.0002
bus	number_wire	64	1.5	0.0015
Processing Unit (8 ReRAM crossbars)				
DAC	resolution	1 bit	4	0.00017
	number	8×128		
ADC	resolution	12	0.0128	0.064
	frequency	5 KHz		
	number	8×32		
ReRAM xbar	number	8	2.4	0.0002
	size	128×128		
	bits per cell	2		
S+A	number	4	0.2	0.0002
S+H	number	8×128	0.01	0.00004
PU Total	number	12	79.4736	0.7758
RAISE Total	PU number	4	466.8344	3.1169

IV. EVALUATION

A. Experimental Setup

Data description. We evaluate the classification accuracy of EEG-based cognition intention using datasets collected from three main BCI paradigms: ERP, SMR, and MRCP. All signals are collected from 64-channel devices. We summarize the details of the dataset used in this work in Table III.

Scheme. We compared the efficiency of the proposed model against (1) DeepConvNet and ShallowConvNet in [6], (2) an LSTM-based model [7], and (3) EEGNet [5]. To evaluate the hardware efficiency, we compared RAISE against CPU and GPU implementations, an FPGA-based implementation [18], and a resistive inference accelerator [12]. We used NVSim [19] to estimate the latency, power and area of ReRAM memory arrays. To simulate the proposed ReRAM accelerator, we use an in-house simulator for system performance evaluation.

B. Classification Accuracy

Within-subject classification. Figure 4 (a) shows the within-subject classification accuracy for models trained on paradigm-specific dataset. For the ERP dataset, there was no significant difference, replicating previous results as reported in [5]. However, for SMR and MRCP, we observed distinct classification performance. LSTM-based models outperform simple convolutional models by averagely 10%, indicating that the complexity of the models greatly improves their ability to capture the information in complex neural activities. EEGNet and the proposed model outperform the LSTM-based model, but there is no statistically significant difference between them.

Cross-subject classification. Figure 4 (b) revealed distinct cross-subject classification accuracy among different BCI paradigms. For the ERP dataset, we see ∼16% accuracy decrease in ShallowConvNet. This is mainly because the ShallowConvNet architecture was specifically designed for oscillatory signal classification [6], thus it may not work well on ERP-based classification tasks. For SMR and MRCP, EEGNet outperforms DeepConvNet, ShallowConvNet, and LSTM-based model by ∼20%. The proposed model in this work, provides 10% more accuracy improvement compared to EEGNet.

Cross-paradigm classification. All previous models demonstrate significant accuracy degradation on cross-paradigm classification as shown in Figure 4 (c). This is because all models are tailored to the distinct characteristics of the expected EEG signals, limiting their application to paradigm-specific tasks. In the contrast, for cross-paradigm classification, the proposed model in this work still retains acceptable classification of 90% on ERP, 73% on SMR, and 82% on MRCP.

TABLE III
SUMMARY OF DATASET.

Paradigm	Source	Channel #	Sample rate	Subject #	Trials/subject	Classe #
ERP	BCI Competition III Dataset II [16]	64	240	2	85	2
SMR	BCI Competition IV Dataset 1 [17]	64	1000	7	26	2
MRCP	PhysioNet [10]	64	512	109	14	4

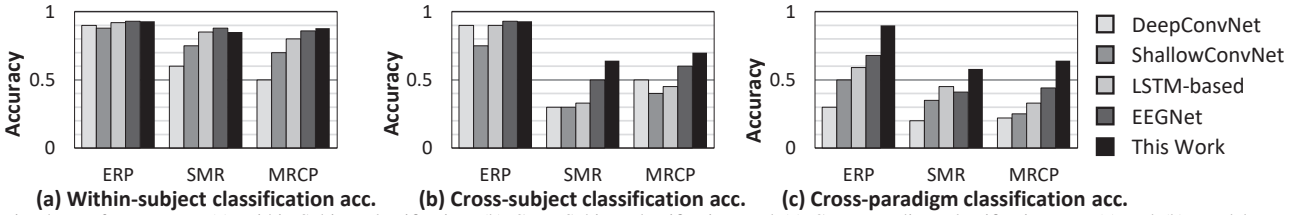

Fig. 4. Performance on (a) Within-Subject classification, (b) Cross-Subject classification, and (c) Cross-Paradigm classification. For (a) and (b), models are trained on single paradigm data; For (c), the parameters in the first two blocks of the model are fixed after training on multi-paradigm data. At the same time, we maintain multiple instances of the third block, each of which is trained on respective paradigm data and selected between different tasks.

C. Hardware Efficiency

Power. Figure 5 (a) shows the power consumption on various platforms. The FPGA-based accelerator provides respectively 62× and 10× power reduction compared with general CPU and GPU platform. ISAAC reduces the power consumption by an order of magnitude compard to FPGA-based design, because its resistive memory based processing engines are more efficient. By tailoring the resistive accelerator design for low-power portable device, RAISE further reduced the power consumption by 2.8×. This improvement mainly attribute to the significantly decreased ADC power, as well as the simplified on-chip communication network.

Inference Performance per Watt. We also compare the energy efficiency by measuring the performance per watt. The performance is evaluates by frame per second (FPS). As shown in Figure 5 (b), the general CPU and GPU platforms both suffer from low computational efficiency. CPU provides only 3 FPS/watt, while GPU provides 15 FPS/watt. The FPGA-based accelerator achieves limited efficiency improvement compared to CPU/GPU with only 27 FPS/watt. In contrast, the resre-sistive accelerators achieve significantly improved computation efficiency. ISAAC achieves 65 FPS/watt. Compared to ISAAC, RAISE improves the performance efficiency by 2.5×.

V. CONCLUSIONS

In this work, we present an algorithm/hardware co-design approach for EEG signal classification in BCI systems. Specifically, we present a compact DNN model which provides better-than-previous classification accuracy in subject-independent tasks across primary BCI paradigms including ERP, SMR, and MRCP. We further propose a cross-paradigm framework through parameter sharing. Based on this model, we implemented a low-power resistive accelerator for processing real-time inference. We demonstrated 2.8× power reduction and 2.5× improvement in performance per watt compared to the state-of-the-art resistive inference accelerator.

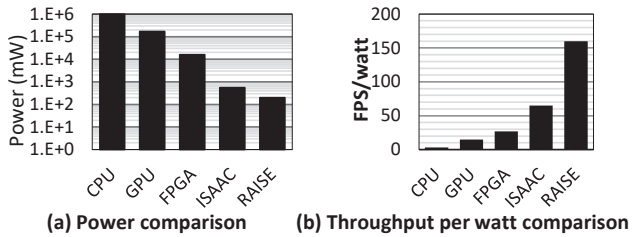

Fig. 5. Comparison on power and throughput per watt.

ACKNOWLEDGMENT

This work was supported by NSF CCF-1725456, NSF 1955246, NSF 1910299, and ARO W911NF-19-2-0107.

REFERENCES

[1] G. Pfurtscheller *et al.*, "Rehabilitation with brain-computer interface systems," *Computer*, 2008.
[2] A. B. Schwartz *et al.*, "Brain-controlled interfaces: movement restoration with neural prosthetics," *Neuron*, 2006.
[3] L.-D. Liao *et al.*, "Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors," *JNER*, 2012.
[4] Q. Zhao *et al.*, "Improving individual identification in security check with an EEG based biometric solution," in *BI*, 2010.
[5] V. J. Lawhern *et al.*, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *JNE*, 2018.
[6] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human brain mapping*, 2017.
[7] D. Zhang *et al.*, "Cascade and parallel convolutional recurrent neural networks on eeg-based intention recognition for brain computer interface," in *AAAI*, 2018.
[8] M. Oquab *et al.*, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in *ICCV*, 2014.
[9] A. Radford *et al.*, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint*, 2015.
[10] "PhysioNet." https://physionet.org/about/database/.
[11] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.
[12] A. Shafiee *et al.*, "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars," in *ISCA*, 2016.
[13] L. Song *et al.*, "PipeLayer: A Pipelined ReRAM-Based Accelerator for Deep Learning," in *HPCA*, 2017.
[14] F. Chen *et al.*, "Regan: A pipelined reram-based accelerator for generative adversarial networks," in *ASP-DAC*, 2018.
[15] M. El Ansary *et al.*, "50nW 5kHz-BW opamp-less ΔΣ impedance analyzer for brain neurochemistry monitoring," in *ISSCC*, 2018.
[16] "BCI Competition III." http://www.bbci.de/competition/iii/.
[17] "BCI Competition IV." http://www.bbci.de/competition/iv/.
[18] Z. Chen *et al.*, "Clink: Compact lstm inference kernel for energy efficient neurofeedback devices," in *ISLPED*, 2018.
[19] X. Dong *et al.*, "NVSim: A Circuit-Level Performance, Energy, and Area Model for Emerging Nonvolatile Memory," in *IEEE TCAD*, 2012.