پیشنهاد پژوهشی

یک پردازنده شتابدهنده یادگیری عمیق برای پردازش سیگنالهای EEG مغز در کاربردهای BCI

DL4BCI: A Specialized Deep Learning Processor for EEG Signal Analysis on BCI-Driven Applications

ارائه شده به: پژوهشکده علوم کامپیوتر، پژوهشگاه دانشهای بنیادی

توسط:

مهدی مدرسی

دانشکده مهندسی برق و کامپیوتر، دانشکده فنی دانشگاه تهران، تهران

modarressi@ut.ac.ir

زمستان ۱۴۰۲

DL4BCI: A Specialized Deep Learning Processor for EEG Signal Analysis on BCI-Driven Applications

1. Introduction

The need for BCI-driven applications. Brain-Computer Interfacing (BCI) aims to translate brain signals into commands that are passed onto an external appliance so as to control it directly just by changing the thought or attention [1]. Among different application, BCI's application in assistive and restorative technologies is gaining more popularity: it facilitates individuals with limited physical abilities interact with computers and devices without using peripheral nervous system and muscles by directly connecting their brain to devices. BCI-driven wheelchairs is an important example of this field that can be considered as the second breakthrough in the wheelchair technology, after the advent of electrical wheelchairs in the mid 1970's. Such device is of extreme help for individuals with damaged neuromuscular pathways or any medical condition severely affecting mobility, including Motor Neuron Disease (MND), Locked-In Syndrome, and Traumatic Spinal Cord Injury (SCI), to facilitate their movement. The considerable incidence rate of these disease highlights the critical importance of assistive devices, including smart wheelchairs, to support the patients. For example, the incidence rate of Amyotrophic Lateral Sclerosis (ALS), as one of the most common types of MND, reaches to 5000 people per year [2]. Further, the global incidence SCI is 10.5 cases per 100,000 persons, with about 330,000 active SCI patients in 2022 in the member states of the Council of Europe [3].

BCI system components. The non-invasive electroencephalography (EEG) method has long been the most common way to extract brain's signals, mainly due to its low cost and overall effectiveness in depicting brain electrical activity. For our specific application, EEG is used to capture signals generated by motor imagery (MI): MI involves mentally rehearsing or imagining a specific movement or action. MI activates many of the same brain regions that are involved in actually executing the physical action, including the primary motor cortex and premotor cortex.

An EEG-based BCI system consists of signal measurement, noise/artefact elimination, feature extraction, classification, and decision-making/control signal generation tasks. The initial step involves signal measurement, which is executed by electrodes on some specific locations on the scalp's surface reading the EEG signal, followed by analogue-to-digital converters (ADCs) which diligently digitize these signals. Subsequently, a processing unit (PU), serving as the vital core of the system, undertakes the responsibility of de-noising the extracted signals, extracting features from the signal, classifying extracted features, and facilitating decision-making.

The superior efficiency and accuracy of the emerging deep learning (DL) in noise removal, feature extraction and classification tasks has made them the principal choice for implementing EEG signal analysis [3]. DL is an emerging branch in the artificial intelligence that uses neural networks to learn how to solve a problem. There is a large body of research on EEG motor imagery signal analysis, as well as on smart DL-based services that can facilitate the use of BCI-driven wheelchairs. A recent survey on the recent advances of DL-based methods in this field can be found in [13]. The Modern DL-based designs extract more signal features, classify signals more accurately, increase tolerance against noise and artefacts more efficiently, implement personalized neural network models that are tailored to an individual's specific EEG patterns and wheelchair control preferences, detect signs of cognitive stress or fatigue from EEG data and adjust system responsiveness or invoke breaks, detect abnormal signals or potentially unsafe navigation scenarios, and implement algorithms that can predict user intentions based on historical data to enable smoother and more intuitive wheelchair control, to name a few.

Research challenge. An important remark about the state-of-the-arts works in the literature is that, unfortunately, the DL research community has paid relatively little attention to the execution speed and energy-efficiency of EEG analysis models due to focusing on accuracy improvement. However, tight energy and resource constraints is a common feature of such medical assistive and health monitoring devices. They are limited to use low-power embedded processors and moderately-sized batteries to keep the battery lifetime, weight, cost, and size of the entire system reasonable. With this limitations, taking the full advantage of complex DL models on wearable devices is quiet challenging: A moderately-sized DL model for EEG

processing requires several hundred thousand to millions of arithmetic operation to process a single time window of the EEG [4]. While a stationary device (such as clinical EEG monitoring devices) can deploy a powerful graphical processing unit (GPU) or microprocessor (CPU) to run such algorithms at real-time, the power consumption (50-200 Watts) and their cost (several hundred dollars) is quite prohibitive for BCI-based wheelchair. In addition, the need for inconspicuous EEG recording (to reduce user's discomfort) leads to electrode placement/count constraints. This, in turn, reduces signal strength and quality. To guarantee the performance on the noisy input data, more sophisticated feature extraction and artefact removal tasks and more robust signal analysis algorithms are needed that makes the already complicated task of EEG analysis even more complicated.
Due to these challenges, a recent survey on BCI-driven wheelchair control systems [2] shows that a very limited number of DL-based methods (out of the excessive volume of the research work) are proposed to be implemented on smart wheelchairs; rather, most existing practical systems replace the more efficient DL with the simpler (but less accurate) alternative methods for EEG analysis. There are some commercial BCI-driven wheelchairs available at the market. Although commercial products rarely reveal the details of the processing unit(s), but their functionalities and the services they offer considerably lag what exists in the state-of-the-art literature.
Methods that combine CNNs (which extract signal features and capture the spatial characteristics of EEG) with RNN (which captures temporal characteristics of EEG) give the best accuracy results and the methods that use a combination of CNNs (for feature extraction) and Transformers have the highest robustness.

The deployment of continuous real-time EEG analysis for BCI-driven wheelchairs is quite challenging. Particularly, several factors contribute to the complexity of this approach, as follows.

- Tight resource constraints for long-term monitoring:
- User discomfort and social stigma: In clinical EEG monitoring, tens of electrodes are connected to some specific locations on the scalp and held fixed to the skin either with a cap or an adhesive. This method has many critical limitations for long-term continuous monitoring, including social stigma and patients discomfort.
- High noise and artefact level: EEG signal is very sensitive to ambient noise, as well as the artefact made by physiological signals that are not of cerebral origin [5][6]. Eye blink, muscle movement, and chewing are some few examples of the EEG artefact sources. This degrades the system robustness. For example, it is reported in [6] that a clinical EEG-based service (for online epileptic seizure detection) makes between 0.1 to 5 false alarms per hour. Being a big concern in clinical settings, noise is even a more severe problem for EEG processing on wearables. First, placing the electrodes in comfort and non-stigmatizing positions reduces the signal quality and strength [6]. Second, continuous EEG monitoring are exposed to daily activities and their associated artefacts.

These challenges impose conflicting requirements: the need for inconspicuous EEG recording leads to electrode placement/count constraints. This, in turn, reduces signal strength and quality. To guarantee the performance on the noisy input data, more sophisticated signal pre-processing and artefact removal tasks and more robust signal analysis algorithms are needed that makes the already complicated task of EEG analysis even more complicated.

The ultimate result of these challenges is increasing the complexity of DL-based EEG analysis task to a degree that do not fit the tight energy budget and compute power of the existing processing units of wearable devices.
Goal. To address the above-mentioned research challenges, we present DL4BCI, an efficient processing unit, as an electronic chip, and its associated software tools for DL-based EEG analysis for BCI-driven wheelchairs. The proposed design aims to enable fast, ultralow-power, accurate, and noise-tolerant execution of a vast range of DL-based EEG analysis tasks in order to facilitate the deployment of DL algorithms for EEG analysis on resource-constrained wearable devices.
The goal will be achieved through the following scientific and technical objectives.

- **Objective 1**. Hardware acceleration of DL-based EEG analysis models. Our primary approach to tackle the excessive compute load and energy consumption of DL-based EEG

analysis models is hardware acceleration. Hardware acceleration of an algorithm refers to designing a special-purpose processor with the internal architecture customized to run that specific algorithm. A hardware accelerator, when designed efficiently, would run a DL model orders of magnitude faster and more energy-efficient than the conventional case, when the software code of the model runs on a general-purpose processor [7].

- **Objective 2**. Domain-specific acceleration. The existing EEG analysis tasks are implemented by one the three well-known DL models, namely multi-layer perceptron (MLP), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformers, or a combination of the four. The DL-based noise removal methods also employ some other deep learning models. Thus, the DL4BCI architecture should be carefully designed as a domain-specific DL-based EEG accelerator, rather than an accelerator for one specific model, to be optimized for all of these models. A domain-specific accelerator is a specialized computing unit that is designed to accelerate a particular domain of applications.
- **Objective 3**. Configuration/execution scheduling generation for the DL4BCI hardware. The processing unit needs a compiler-like tool that takes the target DL model and model hyper-parameters as input, configures the DL4BCI's hardware architecture for the model, and generates a mapping and scheduling to run the model on the hardware.

2. Project Description

DL4BCI's hardware. The proposed hardware is a domain-specific accelerator, which is required to run several DL models efficiently. It consists of a matrix of basic processing unit (BPU) connected by a reconfigurable network-on-chip (NoC). The multiply-and-add (MAC) is the basic arithmetic operation in almost all DL models: it takes place in the form of inner product of two vectors in RNNs and MLPs, in the form of convolution between a sliding two-dimensional filter and an input image in CNNs, and in a combination of these forms in Transformers. Since the computation of the considered DL models uses the MAC operation in different dataflows, BPU implements the MAC operation in an efficient way and then, a reconfigurable NoC is in charge of implementing the right dataflow among BPUs.

In this project, as the starting point of the design, we plan to design the BPU as a datapath that multiplies a vector to a scalar number based on the ΔNN's architecture [8], an efficient DL accelerator we presented for CNN models in a previous work. MAC on two vectors and matrix multiplication can be implemented on this unit by fusing the right number of BPUs.

The NoC is responsible for fusing the BPUs to configure the accelerator for the computation pattern of the target DL model. This involves analysing the communication pattern and dataflow of the target DL models and figuring out how the BPUs should be fused or connected to implement the required computations of the model. Then, a NoC is used to connect the input/output ports of the BPUs in such a way that configure them as a unified larger processing unit customized for the computation pattern of a target DL model. The architecture should also implement DL-based noise removal method, which fortunately, are typically implemented by one or a combination of the aforementioned DL models that we plan to support.

To enable fine-grained architecture reconfiguration, we use the reconfigurable switch-based network we have presented in [9]. In addition to tailoring the DL4BCI's architecture to the running model, since distributing data among the BPUs is done by the NoC, it is the right place to implement memory management methods (such a s data/computation reuse) to reduce memory bandwidth requirement. the repetitive nature of the EEG signals promises the successful deployment of data/computation reuse methods for BCI-driven wheelchairs.

DL4BCI's software. To support the DL4BCI's hardware, we design a compiler-like tool to automate the process of configuring the hardware and scheduling the execution of the model on it. This involves designing an algorithm that takes the model type and hyper parameters (number/size of layers, inputs, outputs, channels) as input and (1) finds the right NoC configuration to adapt DL4BCI's structure to the DL model, (2) partitions the model into several smaller parts, each matching the available computation resources, and (3) schedule the execution of the partitions on the chip. Partitioning and scheduling is necessary since the neural network size is very likely to be larger than the number of BPUs, so the entire model cannot be mapped onto the hardware at once. A typical EEG analysis task would execute a sequence of different DL models, for noise removal, feature extraction, and calcification. Our software tool will calculate the suitable configuration for each model offline. The configurations are loaded onto the architecture at run-time, right before the corresponding model is invoked.

Experimental methodology. To show the effectiveness of DL4BCI, we implement at least two state-of-the-art DL-based methods in the literature for MI-based BCI on it. We plan to use the CNN-RNN MI EEG processing method presented in [10] as one of the case studies. We will keep an eye on new models that will come during the project and will pick at least one more method with state-of-the-art results to our evaluation set. Almost all MI-based BCI methods in the literature are tested and evaluated under the vast set of MI EEG databases [11]. These datasets contain EEG signals of different lengths (form several minutes to several hours) and the MI movement associated with each time window of the signal. We will use the same datasets to evaluate our hardware. This way, the DL4BCI chip will read the EEG of a dataset from a file (to mimic the electrode unit in realistic BCI-driven wheelchair) and run the DL models to generate and display control commands. For artefact and noise removal we will implement the DL-based EEG method presented by L. Benini et. al in [3].
The evaluation metrics are (i) execution time: the time it takes for the processing unit to analyse one second of the input EEG waveform. (ii) Accuracy: The ratio of detected patterns to the all patterns in the input data. (iii) Energy consumption: the total amount of time the processing unit can work with a battery with a specific capacity (e.g. 5000 mah).
Comparison will be done with the results of running the same DL models for BCI on (1) an existing state-of-the-art EEG hardware, (2) a GPU, and (3) an embedded microcontroller.
Project scope and outcome. Through the above case studies, we aim to show that DL4BCI facilities deployment of any MI-based BCI method made by the DL models (that DL4BCI supports) on smart wheelchairs by drastically reducing the energy consumption and execution time of the models. Note that this proposal does not deal with designing and training DL models for BCI: rather, it aims at designing a specialized processor to accelerate the execution of DL-based BCI solutions. DL4BCI's deliverables consist of (i) the implementation of the DL4BCI domain-specific accelerator on the Zynq family of FPGAs, and (ii) its associative software tool in the C++ programming language. The outcome of this project can be used as the main processing unit by smart wheelchair manufacturers.
Note that there are no ethical issues associated with this project, as we use publically available EEG databases at this project.
Project organization. Figure 1 outlines the project steps and the outcome. The project will be structured in three tasks. Each task will be followed by an evaluation and verification step (not shown in the figure) and the step repeats until the required output quality is obtained.

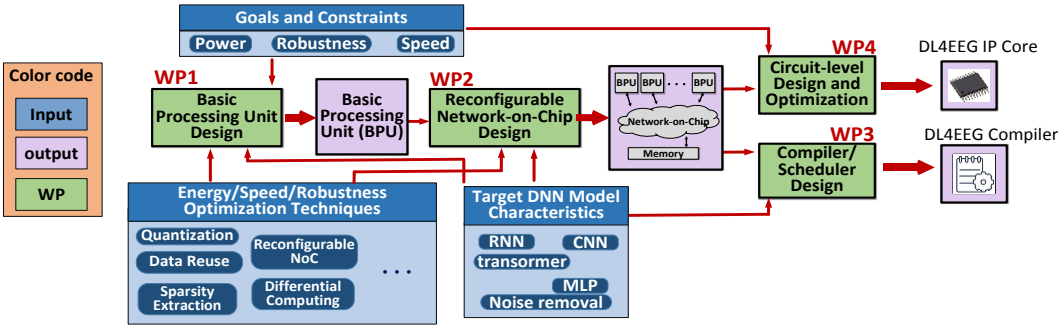

Figure 1. Project steps and outcome

Task1. Basic processing unit (BPU) design
1.1. Identifying the basic processing operations in the supported DL models
1.2. Identifying the basic processing operations in common noise removal algorithms
1.3. Designing an initial architecture for BPU based on the ΔNN core.
1.4. Studying the impact of some promising state-of-the-art optimization techniques on the performance and power usage of BPU and integrating the most effective ones into the design.
Our previous work: We have published more than 30 papers and one book in recent years on hardware acceleration of neural networks. most of the optimization methods that we plan to apply on BPU will be a re-design customized version of the previous works. Particularly, the ΔNN core [8], which uses differential computation to drastically reduce energy usage of deep learning models, promise below 10mW power usage on complex DL models.

Task 2. Network-on-chip (NoC) design
2.1. Studying the target DL computation patterns to list the required BPU fusing schemes

2.2. Memory system design: Finding the right number of memory controller ports and on-chip buffers and their placement in the topology.
2.3. Designing a reconfiguration algorithm to adapt the topology to a given DL model
2.4. Designing methods for data reuse in network-on-chip
Our previous work: The applicants have a strong background in network-on-chip design and optimization and have totally published more than 80 papers in the field. They recently have focused on network-on-chip for DL accelerators that has resulted in publishing 6 joint papers.

Task 3. Neural network compilation and porting
3.1. Design and implement scalable algorithms for scheduling of DL models onto the BPUs.
3.2. Implement a compiler-like tool to automate the process of finding the right network-on-chip configuration for each DL models
Our previous work: We developed a compiler for Coarse Grained Reconfigurable Architecture (CGRAs) which constructs the binary for the complete application (represented in C or Simulink)**.** In DeepMaker project [12], we developed an HLS-based compiler to map DL models to FPGA. Both compilers provide a base for this task.

3. DL4BCI's Significance and Novelties over the State-of-the-arts

There are growing amount of research work on hardware acceleration methods for different deep learning methods, including many methods that target EEG analysis. A comprehensive survey on hardware implementation of DL models can be found a book we have published in 2020 [7]. The scientific novelty of DL4BCI over the existing DL hardware accelerators are twofold. **First**, DL4BCI is a multi-target domain-specific hardware accelerator**.** Based on the application, the best accuracy in EEG analysis is achieved by either RNNs, CNNs, Transforms, or some combination of the three. They also use an MLP (multi-layer perceptron) model for final classification. While there is a large body of research on hardware acceleration of a single model (CNNs, RNNs, MLPs, and Transformers), DL4BCI is capable to configure its internal structure to match the running model. An alternative approach would be designing a separate specialized accelerator architecture for each individual device based on its DL model. DL4BCI, however, has at least two main advantages over this alternative design. First, it is designed once and used in many devices. This effectively reduces the time/effort/cost of designing a separate processor for each device. Second, DL4BCI allows upgrading and changing the DL model or de-noising algorithm, as long as the new algorithm or DL model can be implemented on DL4BCI's cores. We believe these conditions is applicable to many wearable devices, so they can benefit from the reconfiguration capability of DL4BCI. An example of such demands for upgrade in recent years is the emergence of the Transformer DL model that is used in place of the RNN model in some new research work to give higher accuracy and robustness. **Second**, DL4BCI uses the same DL specialized architecture for de-noising and artefact removal. To pursue our goal towards a domain-specific DL-based EEG hardware, the architecture should provide the wearable developers with the ability to implement common de-noising methods with arbitrary parameters. Fortunately, we observed that the common pre-processing methods are constructed by either DL models or signal processing transforms (such as DWT and FFT), which often use the MAC operation as basic compute block. So, the pre-processing can be implemented on the existing hardware resources that are originally designed for DL-based EEG analysis.

References

[1] Värbu, Kaido, Naveed Muhammad, and Yar Muhammad. "Past, present, and future of EEG-based BCI applications." Sensors 22.9 (2022): 3331.
[2]Naser MY, Bhattacharya S. Towards Practical BCI-Driven Wheelchairs: A Systematic Review Study. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2023 Jan 11.
[3x] Towards concerted efforts for treating and curing spinal cord injury, https://assembly.coe.int, 2022.
[3] Arpaia, Pasquale, et al. "How to successfully classify EEG in motor imagery BCI: A metrological analysis of the state of the art." Journal of Neural Engineering 19.3 (2022): 031002.
[4] K. Lee et al., "Real-Time Seizure Detection using EEG: A Comprehensive Comparison of Recent Approaches under a Realistic Setting", in *the Conference on Health, Inference, and Learning*, 2022.
[5] T. M. Ingolfsson, et al., "Energy-Efficient Tree-Based EEG Artifact Detection," IEEE *EMBC*, 2022.
[6] D. Sopic, T. Teijeiro, D. Atienza, A. Aminifar, P. Ryvlin, "Personalized seizure signature: An interpretable approach to false alarm reduction for long-term epileptic seizure detection, Epilepsia, 2022.
[7] M. Daneshtalab and M. Modarressi, Hardware Architectures for Deep Learning. IET publishers, 2020.
[8] H. *Mahdiani, A. Khadem, A. Ghanbari, M. Modarressi, F. Fattahi-Bayat, and M. Daneshtalab, "ΔNN : Power-Efficient Neural Network Acceleration Using Differential Weights," Journal of IEEE Micro, vol. 40, no. 1, 2020.*

[9] A Firuzan, M Modarressi, M Daneshtalab, “A Reconfigurable Network-on-Chip for Efficient Implementation of Neural Networks”, in ReCoSoC Conf., 2015.
[10] Z. Khademi, F. Ebrahimi and H. M. Kordy, "A transfer learning-based CNN and LSTM hybrid deep learning model to classify motor imagery EEG signals", Comput. Biol. Med., vol. 143, Apr. 2022.
[11] Gwon, Daeun, et al. "Review of public motor imagery and execution datasets in brain-computer interfaces." Frontiers in Human Neuroscience 17 (2023): 1134869.
[12] M. Loni, et al., “DeepMaker: A Multi-Objective Optimization Framework for Deep Neural Networks in Embedded Systems”. Microprocessors and Microsystems, 2020.
[13] Arpaia, Pasquale, et al. "How to successfully classify EEG in motor imagery BCI: A metrological analysis of the state of the art." Journal of Neural Engineering 19.3 (2022): 031002.