Original Research Article

Lightweight neural network design for real-time signal processing in brain-computer interfaces

Caiyun Shan, Yanbin Bu

Communication University of China, Nanjing Nanjing, Jiangsu Province, 211172, China

Abstract: Traditional signal processing methods struggle to meet the requirements of real-time performance and accuracy, while deep learning-based approaches, despite their superior performance, suffer from high computational complexity, which hinders real-time processing on embedded devices. To address this challenge, this paper explores the design of lightweight neural networks for real-time signal processing in brain-computer interfaces (BCIs). By analyzing the characteristics of electroencephalogram (EEG) signals, we summarize key considerations for lightweight network design and present an efficient EEG signal classification model using EEGNet as an example. Finally, we discuss the prospects and challenges of lightweight neural networks in BCIs, providing insights for future research in this field.

Keywords: Brain-computer interface; Real-time signal processing; Lightweight neural network; EEGNet

1. Introduction

Lightweight neural networks significantly reduce model parameters and computational costs through structural simplification and computational optimization while maintaining performance, thereby enabling the deployment of deep learning models in resource-constrained scenarios. Integrating lightweight design principles with the real-time processing demands of BCIs offers a promising approach to overcoming the limitations of traditional methods and achieving efficient, accurate EEG signal decoding. This paper focuses on lightweight neural network design strategies tailored for BCIs, aiming to provide new perspectives for advancing research in this domain.

2. Development and current status of lightweight neural networks

2.1. Design motivation and advantages

Although deep learning models have achieved remarkable success in fields such as computer vision and speech recognition, their practical deployment is often hindered by excessive parameters and computational costs. For instance, AlexNet, a classic model, contains 60 million parameters and requires 1 billion floating-point operations (FLOPs), rendering it impractical for real-time execution on mobile or embedded devices. Lightweight neural networks address this issue by minimizing parameter redundancy and computational overhead while preserving performance, thereby enabling efficient execution on devices with limited power, storage, and computational resources[1].

Compared to conventional large-scale networks, lightweight networks offer distinct advantages, including fewer parameters, faster inference speed, and lower resource consumption. For example, MobileNet V1 achieves comparable accuracy to AlexNet on the ImageNet classification task with only 1/30 of its parameters. These advantages make lightweight networks particularly suitable for latency-sensitive and power-constrained applications such as BCIs.

2.2. Representative lightweight network architectures

Numerous lightweight architectures have been proposed and widely adopted in domains like computer vision, offering valuable insights for BCI-oriented designs[2].

SqueezeNet, one of the earliest lightweight CNNs, reduces parameters to 1/50 of AlexNet while maintaining comparable accuracy. Its core innovation lies in the "Fire module," which combines a squeeze layer (using 1×1 convolutions for channel compression) and an expand layer (employing parallel 1×1 and 3×3 convolutions to enhance feature representation). Additionally, delayed downsampling strategies are utilized to preserve spatial information in early layers.

MobileNet Series: Designed for mobile deployment, MobileNet employs depthwise separable convolution, decomposing standard convolution into depthwise (channel-wise) and pointwise (1×1) convolutions. MobileNet V2 introduces inverted residual blocks with linear bottlenecks to improve parameter efficiency, while MobileNet V3 leverages neural architecture search (NAS) and the h-swish activation function to optimize the trade-off between speed and accuracy[3].

ShuffleNet: This architecture enhances inter-group information flow through group convolution and channel shuffle operations. By partitioning input channels into groups and shuffling them after group-wise processing, ShuffleNet reduces computational costs while maintaining accuracy. Pointwise group convolution further minimizes parameters, enabling higher classification accuracy than MobileNet under similar computational budgets.

Key principles of lightweight network design include:

1. **Efficient convolution operations** (e.g., depthwise, group, and pointwise convolutions);
2. **Compact network topologies** (e.g., smaller kernels, reduced downsampling layers);
3. **Modularity and reuse** (e.g., Fire modules in SqueezeNet);
4. **Dimensionality reduction techniques** (e.g., channel compression, low-rank decomposition).

3. Lightweight network design for BCIs

3.1. Design considerations and challenges

Direct application of lightweight CNNs to EEG signal processing faces limitations due to the unique characteristics of EEG data: high-dimensional non-stationary time series, low signal-to-noise ratios, and significant inter-subject variability. Additional challenges include limited sample sizes and high costs associated with data acquisition.

Design Strategies:

Time-frequency representation: Integrate spectral modeling (e.g., short-time Fourier transform (STFT) layers or wavelet transforms) with spatiotemporal convolutions to capture joint time-frequency features.

Spatial correlations: Model dynamic inter-electrode relationships using attention mechanisms or graph convolutional networks (GCNs) to emphasize discriminative brain regions.

Regularization and data augmentation: Mitigate overfitting via L2 regularization, dropout, generative adversarial networks (GANs), or channel mixing techniques.

Performance trade-offs: Optimize models based on application requirements—prioritize real-time inference for control tasks (e.g., prosthetics) or accuracy for diagnostic applications (e.g., emotion recognition).

3.2. Enhanced depthwise separable convolution

Standard depthwise separable convolution processes temporal signals but overlooks spectral information. We propose **frequency-aware depthwise convolution**, which applies depthwise operations to time-frequency representations (e.g., STFT-transformed EEG signals) to capture spectral-temporal patterns. Inspired by ShuffleNet, **grouped pointwise convolution** partitions channels into spatial groups (e.g., by brain regions) and shuffles them to reduce parameters while preserving spatial structure.

3.3. Attention mechanisms

Efficient attention variants, such as Squeeze-and-Excitation (SE) and Gather-Excite (GE) modules, adaptively weight channels or spatial regions with minimal additional parameters. For EEG processing, **topology-aware attention** incorporates electrode spatial relationships via graph convolutions, while **dynamic attention** models time-varying feature importance through gated recurrent units (GRUs) or self-attention mechanisms.

3.4. Cross-layer feature fusion

Multi-scale feature fusion enhances representation completeness. Inspired by feature pyramid networks (FPNs), we propose **hierarchical fusion** using 1×1 convolutions to compress shallow features before combining them with deeper abstractions. **Multi-resolution temporal pathways** process EEG signals at varying time windows and sampling rates to capture delta, theta, and alpha rhythms, mimicking the cross-scale time-frequency modeling of MSMF-Net.

3.5. Few-Shot incremental learning

To address inter-subject variability and limited data, **meta-learning frameworks** (e.g., Model-Agnostic Meta-Learning, MAML) and **knowledge distillation** techniques enable rapid adaptation to new users with minimal retraining. Gradient-based sample selection strategies prioritize high-impact samples to further improve update efficiency.

4. Experiments and results

We validate our design by enhancing EEGNet with frequency-aware depthwise convolution, grouped pointwise convolution, and attention mechanisms, proposing **EEGNet-Lite**. Evaluations on the BCI Competition IV 2b dataset (motor imagery classification) yield the results shown in **Table 1**.

Table 1. Performance comparison on BCI competition IV 2b dataset.

Model	Parameters (M)	FLOPs (M)	Accuracy (%)	Inference Time per Sample (ms)
EEGNet	0.87	172	71.6	3.2
ShallowConvNet	0.21	48	67.4	1.8
DeepConvNet	3.26	702	70.1	13.5
EEGNet-Lite (ours)	0.32	58	73.2	1.4

EEGNet-Lite achieves higher accuracy (73.2%) with fewer parameters (0.32M) and lower computational costs (58M FLOPs), demonstrating superior efficiency. Ablation studies reveal that frequency-aware convolution contributes most to performance improvement, followed by attention mechanisms and grouped pointwise convolution.

5. Conclusion and outlook

This paper explores lightweight neural network design for real-time signal processing in BCIs. By integrating frequency-aware convolutions, dynamic attention mechanisms, and cross-scale feature fusion, EEGNet-Lite achieves state-of-the-art performance with minimal resource demands. Future work will focus on adaptive architectures for cross-subject generalization and hardware-software co-design to further advance practical BCI deployment.

About the author

Caiyun Shan, born in December 1980, female, originally from Funing County, Jiangsu Province, Han ethnicity, education: Bachelor's degree, professional title: Associate Researcher, research direction: Computer Science and Technology.

Yanbin Bu, borning in 1984, was born in Guanyun County, Lianyungang City, Jiangsu Province. He holds a master's degree and is a university professor. His research direction is artificial intelligence.

References

[1] He, Q., Wu, H., Tian, F., et al. A real-time brain-computer interface system for controlling short message transmission. Chinese Journal of Medical Physics, 2012, 29(3): 3386-3392.

[2] Zhu, X., Han, Y., Hao, Y., et al. Real-time spike sorting algorithm based on probabilistic neural network. Journal of South China University of Technology (Natural Science Edition), 2012, 40(6): 48-55.

[3] Xie, Y., Liu, Y., Chen, C., et al. An embedded acceleration method for lightweight convolutional neural networks. Journal of Chinese Computer Systems, 2023, 44(7): 1345-1351.

[4] Zhang, T., Xu, D., Chen, J., et al. Patent main path analysis in brain-computer interface field. China Medical Devices, 2025, 40(3): 8-15.