DARNet: Dual Attention Refinement Network with Spatiotemporal Construction for Auditory Attention Detection

Sheng Yan, Cunhang Fan*, Hongyu Zhang, Xiaoke Yang,
Jianhua Tao and Zhao Lv
Anhui Province Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology,
Anhui University, Hefei, 230601, China;
*Corresponding Author Email: cunhang.fan@ahu.edu.cn
Code Address: https://github.com/fchest/DARNet.git

NEURAL INFORMATION PROCESSING SYSTEMS

Contributions

- **A Novel Auditory Attention Decoding Architecture**: We proposed a novel architecture for AAD, which could fully leverage the spatiotemporal features and capture long-range latent dependencies of EEG signals.
- **Optimized Decoding Performance**: The DARNet shows substantial improvement across all datasets and achieves further parameter reduction, requiring only 0.08 million parameters.

Background

Cocktail Party Phenomenon

The "cocktail party phenomenon" refers to the ability of individuals with normal hearing to focus on a specific sound, such as a speaker's voice, in a noisy, multi-speaker environment. This selective auditory attention allows people to isolate relevant sounds from complex acoustic scenes, playing a crucial role in communication, especially in noisy settings.

Auditory Attention in Hearing Impairment

For individuals with hearing impairments, focusing on specific sounds in noisy environments is challenging. This reduced ability to process auditory signals, known as impaired auditory attention, leads to communication barriers in social and professional settings, causing frustration, social isolation, and even cognitive decline. Improving auditory attention mechanisms for hearing-impaired individuals has become a key research focus to address these challenges.

Auditory Attention Detection (AAD)

AAD aims to decode the neural mechanisms underlying auditory attention by analyzing brain activity. Techniques such as EEG, MEG, and ECoG have been used to study how the brain processes and focuses on sound. Among these, EEG is particularly suited for AAD due to its non-invasive nature, high temporal resolution, and practicality for real-world applications, making it ideal for developing assistive technologies for individuals with hearing impairments.

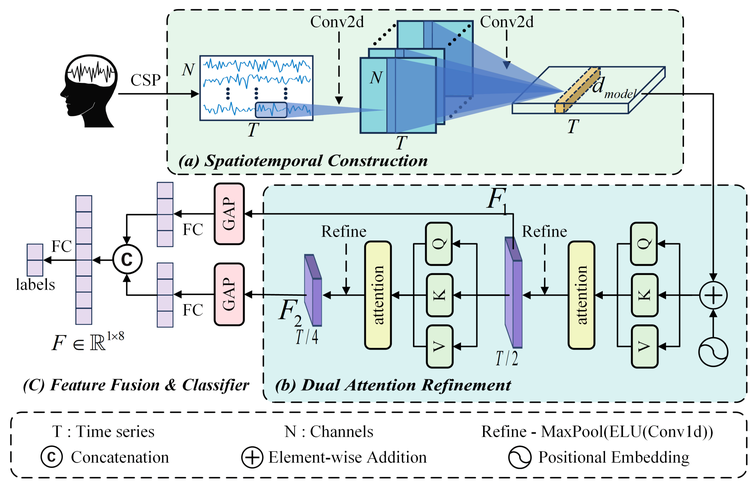
Limitations

Neglecting the Spatial Feature Distribution

EEG signals reflect both temporal and spatial patterns of brain activity, yet previous AAD methods have primarily focused on local temporal features, neglecting the spatial distribution of EEG data. This limits the ability to fully capture the brain's response to auditory stimuli.

Difficulty in Capturing Long-Range Dependencies

Human attention is dynamic, with prior brain activity influencing subsequent responses. However, existing AAD models struggle to capture long-range temporal dependencies due to model depth and noise in EEG data, hindering their ability to accurately decode auditory attention.



Our Proposed DARNet Method

To address the limitations of previous AAD methods, we introduce DARNet, a model that captures the spatiotemporal and long-range dependencies in EEG signals to improve auditory attention detection.

Spatiotemporal Construction Module

This module captures both temporal and spatial distribution features of EEG signals. Temporal convolutional layers are first applied to capture dynamic temporal patterns in EEG signals, forming a temporal feature set. A spatial convolutional layer is then used to analyze the spatial relationships across EEG channels, providing a comprehensive spatiotemporal representation of the EEG data that enhances understanding of the brain's responses to auditory stimuli.

Dual Attention Refinement Module

The dual attention module captures long-range dependencies in EEG signals to model the dynamic nature of auditory attention. This module applies a self-attention refinement mechanism, reducing noise and outliers by compressing the EEG series. Stacking two self-attention layers allows the model to capture multiple levels of temporal dependencies, effectively enriching temporal features and enhancing model robustness.

Feature Fusion & Classifier Module

This module combines features from multiple layers to preserve essential discriminative information while reducing redundancy. Outputs from each attention layer are projected to a common dimension, concatenated, and fed into a fully connected layer for final auditory attention prediction. This design allows the classifier to leverage the rich spatiotemporal features, improving decoding accuracy and model performance.

Experimental Results

- **Datasets:** we conduct experiments on three publicly available datasets, KUL, DTU and MM-AAD, which are commonly used in auditory attention detection to evaluate the effectiveness of our DARNet. KUL and DTU only contain EEG data of the auditory stimulus scenes. MM-AAD contains EEG data of the audio-only scene and the audio-visual scene.
- **Comparative Study:** "-" means there are no corresponding experiments conducted or no results in the corresponding paper. The results annotated by * are taken from DBPNet.

Dataset	Scene	Model	Decision Window		
			0.1-second	1-second	2-second
KUL	audio-only	SSF-CNN* [37]	76.3 ± 8.47	84.4 ± 8.67	87.8 ± 7.87
		MBSSFCC* [15]	79.0 ± 7.34	86.5 ± 7.16	89.5 ± 6.74
		BSAnet [38]	-	93.7 ± 4.02	95.2 ± 3.08
		DenseNet-3D [39]	-	94.3 ± 4.3	95.9 ± 4.3
		DBPNet* [20]	87.1 ± 6.55	95.0 ± 4.16	96.5 ± 3.50
		DARNet(ours)	**91.6 ± 4.83**	**96.2 ± 3.04**	**97.2 ± 2.50**
DTU	audio-only	SSF-CNN* [37]	62.5 ± 3.40	69.8 ± 5.12	73.3 ± 6.21
		MBSSFCC* [15]	66.9 ± 5.00	75.6 ± 6.55	78.7 ± 6.75
		BSAnet [38]	-	83.1 ± 6.75	85.6 ± 6.47
		EEG-Graph Net [40]	72.5 ± 7.41	78.7 ± 6.47	79.4 ± 7.16
		DBPNet* [20]	75.1 ± 4.87	83.9 ± 5.95	86.5 ± 5.34
		DARNet(ours)	**79.5 ± 5.84**	**87.8 ± 6.02**	**89.9 ± 5.03**
MM-AAD	audio-only	SSF-CNN* [37]	56.5 ± 5.71	57.0 ± 6.55	57.9 ± 7.47
		MBSSFCC* [15]	75.3 ± 9.27	76.5 ± 9.90	77.0 ± 9.92
		DBPNet* [20]	91.4 ± 4.63	92.0 ± 5.42	92.5 ± 4.59
		DARNet(ours)	**94.9 ± 4.79**	**96.0 ± 4.00**	**96.5 ± 3.59**
	audio-visual	SSF-CNN* [37]	56.6 ± 3.82	57.2 ± 5.59	58.2 ± 6.39
		MBSSFCC* [15]	77.2 ± 9.01	78.1 ± 10.1	78.4 ± 9.57
		DBPNet* [20]	92.1 ± 4.47	92.8 ± 5.94	93.4 ± 4.86
		DARNet(ours)	**95.8 ± 4.04**	**96.4 ± 3.72**	**96.8 ± 3.44**

Reference

[1] Ni Q, Zhang H, Fan C, et al. DBPNet: Dual-Branch Parallel Network with Temporal-Frequency Fusion for Auditory Attention Detection[C]//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24). 2024.

[2] Siqi Cai, Tanja Schultz, and Haizhou Li. Brain topology modeling with eeg-graphs for auditory spatial attention detection. IEEE Transactions on Biomedical Engineering, 2023.

[3] Cai S, Schultz T, Li H. Brain topology modeling with EEG-graphs for auditory spatial attention detection[J]. IEEE Transactions on Biomedical Engineering, 2023.