A Deep Channel Attention Transformer for Multimodal EEG-EOG-Based Vigilance Estimation

Jiahui Pan (panjiahui@m.scnu.edu.cn)
School of Software, South China Normal University, Guangzhou, 510631, China

Dehua Lu (2023024275@m.scnu.edu.cn)
School of Software, South China Normal University, Guangzhou, 510631, China

Abstract

An accurate estimation of driver vigilance is crucial for reducing fatigue-related incidents and traffic accidents. Despite advances in the field of fatigue detection, effective utilization of multimodal information remains a major challenge. Additionally, prevalent methodologies predominantly focus on local features, overlooking the importance of global features in this context. To solve the above problems, we propose the deep channel attention transformer (DCAT) model, which can effectively utilize multimodal information and extract local-global features for fatigue detection regression tasks. We first introduce a novel multimodal approach that integrates electroencephalography (EEG) and electrooculogram (EOG) data, capitalizing on their complementary strengths to enhance the understanding and assessment of fatigue states. Then, the DCAT model utilizes multimodal information by extracting local and global features using channel attention and transformer encoder modules, respectively. Our evaluation of the SEED-VIG and SADT public datasets showcases the model's superior performance compared to that of the state-of-the-art baselines.

Keywords: Electroencephalography (EEG); Electrooculogram (EOG); Driver vigilance estimation; Transformer; Channel attention.

Introduction

Gobal concerns about public safety have intensified with the increasing incidence of traffic accidents, which are predominantly attributed to driver fatigue. With the increase in automobile usage and driving intensity, fatigued driving has emerged as a significant cause of accidents. Therefore, the detection and prevention of fatigued driving are crucial tasks in traffic safety research.

Fatigue detection methods are generally categorized into subjective assessment, behavioral assessment, and physical testing methods. While subjective assessments provide valuable tools, their accuracy is limited by individual psychological factors. Behavioral assessments, which involve monitoring facial expressions and body postures, are subject to environmental interference. Consequently, there has been growing interest in physiological indicators, such as electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) signals, for assessing human body states. These methods offer comprehensive health information and accurately reflect fatigue levels, thus holding a promising future in fatigue assessment. Nevertheless, several challenges remain unresolved:

Multimodal information integration has not yet been optimally leveraged for its full potential. A substantial body of work has focused on unimodal detection using various physiological signals, including EEGs, EOGs, and EMGs. Y. Zhang et al. (2022) introduced an automatic weighting variable to adaptively and quantitatively assess the significance of different feature dimensions. This approach effectively addresses the challenge of limited EEG training samples. G. Zhang and Etemad (2023) distilled EEG representations using knowledge distillation via capsule based architectures and used it for various tasks including fatigue detection with good results. Nevertheless, as illuminated by the findings of (Zheng & Lu, 2017; Pan et al., 2023), the synergy of different physiological signals is crucial, because they provide complementary information. Consequently, harnessing the power of multimodal information efficiently emerges as a pivotal challenge.

The local-global features of multimodal information deserve more attention. The need for global feature extraction is particularly crucial given the intricate and interconnected nature of brain functions. The utilization of advanced analytical techniques, such as neural networks, plays a pivotal role in identifying and interpreting these global features. The self-attention mechanism is adaptable at handling sequential data (Vaswani et al., 2017). Its capacity to weigh and integrate information from various parts of the input sequence allows for a more holistic and comprehensive representation of the global features. This capability aligns well with the complexity of brain signals, where the interplay of different neural activities and their collective influence on fatigue states is essential for accurate detection and analysis. In summary, this approach proves advantageous when dealing with challenges.

To address these issues, we propose the deep channel attention transformer (DCAT) model for multimodal EEG and EOG inputs. We first perform a simple fusion of the multimodal data, followed by DCAT to accomplish effective extraction of local-global features and for fatigue detection regression tasks. In summary, our contributions include the following:

1. By integrating EEG and EOG data, we developed a novel multimodal fatigue detection paradigm. This approach, leveraging the complementary strengths of EEG and EOG, offers a comprehensive methodology for identifying of fatigue states.
2. We propose an advanced DCAT model to enhance the fa-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

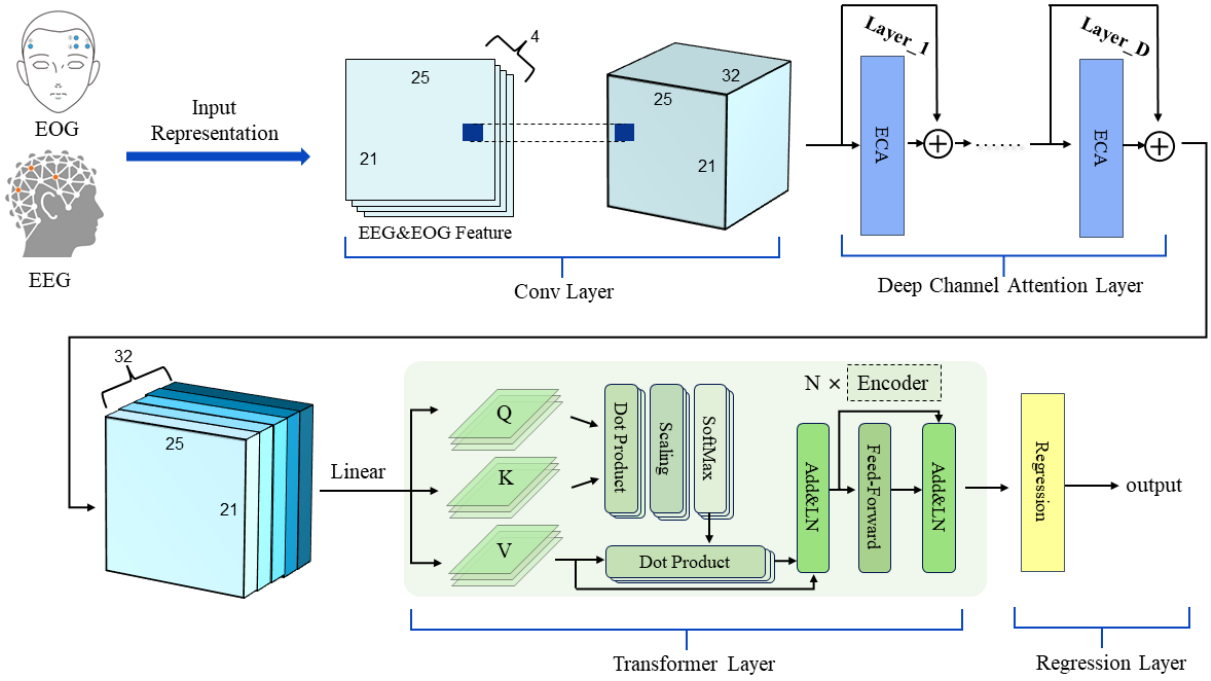

Figure 1: The complete architecture of the paradigm, which consists mainly of input representation and DCAT. The input representation is responsible for extracting features from the raw data and DCAT is responsible for fatigue detection.

tigue detection accuracy. This model excels in extracting global features from multimodal inputs, significantly improving performance and showing potential for fatigue assessment, thus advancing multimodal data processing techniques.

3. We conducted extensive evaluations using the SEED-VIG and SADT public datasets. The experimental results show that our model achieves state-of-the-art performance in fatigue detection, proving its practical applicability in real-world scenarios.

Method

Architecture Overview

Figure 1 shows an introduction to the whole structure, including the Input Representation, which handles the raw data for both modalities, and the DCAT model for fatigue detection. Input representation is used to complete the initial feature extraction of multimodal information. DCAT is subdivided into four parts: the Conv layer, the deep channel attention (DCA) layer, the Transformer layer, and the regression layer. Specifically, in the Conv layer, the 2D-CNN is used to extract preliminary shallow features. The shallow features are input into the next layer of the DCA to efficiently extract deep local features. Then Multi-Head Self-Attention in Transformer is used to extract the neglected global features. The extraction of global features can improve the robustness of the model. In the final regression layer, some processing was performed and linear layers were used to accomplish the regression task.

Input Representation

This module is pivotal for transforming raw EEG and EOG signals into pertinent features that enhance the predictive accuracy of the model. The module encompasses three primary processes: Data Processing, Feature Extraction, and Feature Fusion.

Data processing: To mitigate artifact and reduce computational effort, both EEG and EOG signals are initially downsampled from 1000 Hz to 200 Hz. The data were further refined using bandpass filtering (1-50 Hz) and bandstop filtering (49-51 Hz). In this study, we also extracted full-band features with a 2 Hz frequency resolution for comprehensive analysis.

Feature Extraction: Building upon prior research(Shi, Jiao, & Lu, 2013; Duan, Zhu, & Lu, 2013; Pan et al., 2023), we utilized power spectral density (PSD) and differential entropy (DE) as our primary features, with an 8-second non-overlapping window, due to their proven effectiveness in alertness detection and emotion recognition within brain-computer interfaces. For enhanced feature refinement, techniques such as the moving average (MA) and linear dynamic system (LDS) are applied to these segments, resulting in four distinct feature sets: PSD-MA, DE-MA, PSD-LDS, and DE-LDS.

Feature Fusion: For EEG and EOG with correlated and complementary information, early fusion methods involving direct and simple splicing are more robust than late fusion methods are. Therefore, we fused the EEG and EOG signals into a mixed matrix:

$$X_i^{fused} = \left\{ x_i^{eeg} \cup x_i^{eog} : \forall i \in [1, N] \right\} \tag{1}$$

where N denotes the number of subjects and i denotes the i-th subject. Accordingly, we have

$$fused_l\left(f_t, ch, w_f\right) : \forall l \in [1, L] \tag{2}$$

where $fused_l$ denotes a sample after fusion, f_t denotes the number of features we extracted, and ch denotes the total number of channels, and w_f denotes the 25 divided frequency bands.

Deep Channel Attention Transformer

1) Conv layer: We first use a convolutional layer to extract the original features, map the features into a different feature space and extract the spatial features with a convolutional kernel in the convolutional layer to learn useful information from the data. Specifically, the fusion feature $fused_l\left[f_t, ch, w_f\right]$ passes through a 2D-CNN with a step size of s and a padding of p, followed by a rectified linear unit (ReLU) layer.

2) Deep Channel Attention layer: In this layer, we use multi-layer efficient channel attention (ECA) (Wang et al., 2020) to extract deeper localized features. The specific structure of the ECA is shown in Figure 2. ECA represents an innovative advancement in convolutional neural network architectures and is designed to enhance feature representation through a dynamic channelwise attention mechanism.

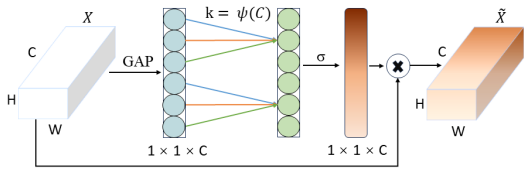

Figure 2: Diagram of the efficient channel attention (ECA) module. GAP stands for global average pooling.

The process begins with a step that compresses the spatial dimensions while retaining critical information across different channels. For the specific implementation we used global average pooling (GAP). This step allows for fast computation of features along the channel dimensions while avoiding dimensionality reduction. The next section utilizes channel attention to capture local cross-channel interactions, which is intended to be both efficient and effective. Given the aggregated feature y, channel attention can be learned by

$$\omega = \sigma(\mathrm{W}y) \tag{3}$$

The ECA module employs a band matrix W_k to learn the channel attention:

$$\begin{bmatrix} w^{1,1} & \dots & w^{1,k} & 0 & 0 & \dots & \dots & 0 \\ 0 & w^{2,2} & \dots & w^{2,k} & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & w^{C,C-k+1} & \dots & w^{C,C} \end{bmatrix}$$

W_k involves $k \times C$ parameters, where k denotes the number of neighbors and C denotes the number of channels. By doing so, the weights of y_i will consider only his k neighbors, and we can obtain the following result:

$$\omega_i = \sigma(\sum_{j=1}^{k} w_i^j y_i^j), y_i^j \in \Omega_i^k \tag{4}$$

where Ω_i^k indicates the set of k adjacent channels of y_i. In the specific experiments, we used a one-dimensional convolution for the implementations, with k representing the size of the convolution kernel. To avoid unnecessary computation, the size of the convolution kernel depends on the channel dimension C.The kernel size k can be adaptively determined by

$$k = \psi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{5}$$

where $|t|_{odd}$ indicates the nearest odd number of t. In the experiments, we set y and b to 2 and 1, respectively. Subsequent to convolution, the extracted features undergo a nonlinear transformation through an activation function, normalizing them for the next step. The final step involves scaling the original input by these normalized features, thereby emphasizing more informative elements and suppressing fewer relevant ones. This design enhances not only the representational capacity of the network but also the computational demand. This results in significant improvements in learning effectiveness without substantial computational overhead. Between each ECA layer, we also use a residual network(He et al., 2016) to prevent gradual explosions.

The core of the ECA mechanism is to assign different importance to the features of different channels, which can capture local attention across channels without dimensionality reduction, in contrast to the traditional attention mechanism (Hu, Shen, & Sun, 2018), which reduces the number of parameters and computational complexity. This attention mechanism can be further enhanced by multi-layer ECA, where each layer can adjust and optimize the feature representation at different levels, allowing the model to perform better in extracting deeper and more complex features.

3) Transformer layer: In previous research, global features have often been overlooked. Given the presence of global correlation in biosignals, extracting global features of multimodal information is extremely beneficial for improving the robustness of the model. In this module, we choose an encoder module from the transformer to extract global features, compensating for the limited receptive field of the convolutional module. The output of the preceding module is linearly transformed and subsequently fed into the Q, K, and

V (query, key, value) components. The encoder architecture, leverages a dot product operation to ascertain the correlation among various tokens. This is accomplished by applying the dot product between the Q and K matrices. To mitigate the risk of vanishing gradients and ensure a stable training process, a scaling factor is judiciously introduced. Subsequently, the resultant matrix undergoes normalization through a Softmax function, which effectively generates a weighting matrix, commonly referred to as the attention score. This attention score is then strategically applied to the V matrix through another dot product, culminating in a weighted representation that embodies the focused attention mechanism of the model. In the case of self-attention, the contents of Q, K, and V remain consistent.The computation is as follows:

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

The term d_k represents the length of a token. In addition, we also employ a multihead attention (MHA) mechanism to further enhance the diversity of features. In this mechanism, tokens are evenly divided into h parts. Each part is individually fed into a self-attention module for computation. Finally, the results from each part are concatenated to obtain the final output. The entire process can be represented as follows:

$$\begin{aligned} MHA(Q,K,V) &= [head_0;...;head_{h-1}], \\ head_l &= Attention(Q_l,K_l,V_l) \end{aligned} \tag{7}$$

where Q_l, K_l, and V_l represent the query, key, and value matrices, respectively, for the l-th attention head. In addition, the fitting ability of the model was enhanced. The residual connection followed by layer normalization (LN) occurred between the feedforward layer and the MHA layer. The input and output dimensions of this module remain unchanged. Within the entire Transformer layer, the encoder structure is repeated N times.

4) regression layer: This layer transforms features into a continuous output value confined between 0 and 1. It commences with a LayerNorm module, which normalizes the input across its features. Following normalization, an AvgPool layer effectively reduces the data dimensionality while preserving essential information, and adapting to varying input sizes. Subsequently, the data are subjected to a linear transformation via the linear layer, mapping it to a singular output value per instance, in line with the regression objective. This process concludes with the application of the tanh activation function and further adjustments to move and scale the output range to [0, 1], thus ensuring that the output of the layer is best suited for regression tasks that require bounded continuous prediction.

Experiment and Results

Datasets

To evaluate our proposed model, we selected the SEED-VIG (Zheng & Lu, 2017) and SADT (Cao et al., 2019) datasets for testing. The SEED-VIG dataset consists of 23 participants. The EEG and EOG signals were sampled at a rate of 1000 Hz. The EEG signals were recorded using 18 electrodes (including a reference electrode), while the EOG signals were recorded using 4 electrodes. The sustained attention driving task (SADT) dataset consists of 27 participants, and the experiment lasts for 60-90 minutes, during which only EEG data are recorded.

Experiment Details

Due to the disparity in the data range between EEG and EOG data, we employed a normalization technique to control for their distributions. The construction and training of DCAT was performed on an NVIDIA 4060Ti, utilizing Python 3.8 and PyTorch. All the data are divided into the training set and the test set according to the ratio of 8:2, and five cross-validation methods are carried out. We employed the AdamW optimizer to effectively train our model.

Evaluation Method

To evaluate the effectiveness of the DCAT model, we employed two metrics: the root mean square error (RMSE) and the Pearson correlation coefficient (PCC). The RMSE is calculated as follows:

$$RMSE(Y,\hat{Y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{8}$$

where y_i represents the actual values and $\hat{y}_i$ represents the predicted values.

The PCC is calculated as follows:

$$PCC(Y,\hat{Y}) = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{9}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, $\bar{y}$ represents the mean of the actual values, and $\bar{\hat{y}}$ represents the mean of the predicted values. In summary, our objective is to train the model to achieve lower RMSE values and higher PCC values.

Comparison Method

In the development and assessment of our DCAT model, a key focus has been its comparative analysis against a range of state-of-the-art methods and established baseline models. We compare the DCAT with the following baselines:

- (G. Zhang & Etemad, 2023):A novel knowledge distillation pipeline is currently available for use in capsule-based architectures for distilling EEG representations.
- (Pan et al., 2023):A new multimodal detection method for estimating driver vigilance is introduced, incorporating residual attention blocks and a capsule attention mechanism.
- (Ding, Zhang, & Eskandarian, 2022):EEG-Fest, a novel solution, is presented as a generalized few-shot model designed to address existing limitations.

Table 1: Comparison of the performance results of different methods on the SEED-VIG and SADT datasets

Paper	Method	Dataset	RMSE	PCC
(Jiang et al., 2020)	O-MV-T-TSK-FS	SADT	0.2+	-
(G. Zhang & Etemad, 2021)	LSTM-CapsAtt	SEED-VIG	0.029	0.989
(Song et al., 2021)	DCRA_E	SEED-VIG	0.035	0.980
	DCRA_M		0.023	0.985
(Y. Zhang et al., 2022)	AWIRVFL	SEED-VIG	0.063	-
		SADT	0.108	
(Ding et al., 2022)	EEG-Fest	SEED-VIG	0.030	0.980
(G. Zhang & Etemad, 2023)	Distillation	SEED-VIG	0.025	0.993
(Pan et al., 2023)	Res-att-capsnet	SEED-VIG	**0.016**	-
		SADT	0.108	
Ours	**DCAT**	SEED-VIG	0.018	**0.998**
		SADT	0.128	-

- (Y. Zhang et al., 2022):An auto-weighting incremental random vector functional link (AWIRVFL) network model was proposed for EEG-based driving fatigue detection.
- (G. Zhang & Etemad, 2021):An architecture consisting of a deep long short-term memory (LSTM) network followed by a capsule attention mechanism is described.
- (Song, Zhou, & Wang, 2021): This model employs a coupling layer to connect two single-modal autoencoders, constructing a joint objective loss function optimization model which comprises single-modal loss and multi-modal loss.
- (Jiang et al., 2020):An online multi-view and transfer TSK fuzzy system for driver drowsiness estimation is proposed, utilizing the 1st-order TSK fuzzy system and integrating the nature of multi-view settings into the existing transfer learning framework.

Comparison with State-of-the-Art Methods

In our comprehensive analysis, the DCAT model was evaluated against various state-of-the-art methods on the SEED-VIG and SADT datasets. In regard to the SEED-VIG dataset, we adopted a 5-fold cross-validation method to evaluate the model's performance, mirroring the approach in other studies. The results are presented in Table 1. As shown in the results, DCAT demonstrated a remarkable RMSE of 0.018 and a PCC of 0.998. This performance notably surpasses that of established methods such as Distillation and AWIRVFL. Res-att-capsnet achieved optimal performance on two datasets, which may be attributed to experimental methodologies distinct from ours. Specifically, the application of ten-fold cross-validation enabled the extraction of a richer set of features. This performance indicates that the whole model can effectively utilize effective channel information to enhance representation learning. Experimental results on the SADT dataset are also presented in Table 1, where the DCAT model demonstrates exceptional performance, confirming its robustness and effectiveness. The results from this dataset further emphasize the significant contributions of the channel attention and Transformer mechanisms within DCAT, as seen in the marked improvements over the baselines and other SOTA methods.

Table 2: Comparison of our architecture with unimodal features and multimodal features

Modality	RMSE±SD	PCC±SD
EEG	0.0213±0.0017	0.9967±0.0005
EOG	0.0296±0.0021	0.9936±0.0009
EEG+EOG	**0.0180±0.0023**	**0.9977±0.0007**

Through this comparative analysis, it is evident that the integration of deep learning with channel attention mechanisms in a Transformer framework, as employed in DCAT, not only enhances the model's performance but also contributes significantly to its ability to accurately and reliably detect fatigue states from EEG and EOG data.

Ablation Experiment

Modality

To address the current challenge that multimodality cannot be effectively utilized, we propose an approach to feature fusion using early fusion. In this study, we employed two modalities, EEG and EOG, for fatigue detection. EEG reveals changes in electrical signal activity in the cerebral cortex, especially within specific frequency bands. The EOG captures variations in muscle activity in the frontal region, particularly

Table 3: Ablation Study of the DCA and Transformer

DCA	Transformer	Dataset	RMSE
✓		SEED-VIG	0.0312
	✓	SEED-VIG	0.0231
✓	✓	SEED-VIG	**0.0180**
✓		SADT	0.1390
	✓	SADT	0.1348
✓	✓	SADT	**0.1280**

blinking and saccades, which are closely related to fatigue. We performed unimodal and multimodal tests on the SEED-VIG dataset, and the results are shown in Table 2. According to Table 2, multimodal features are superior to unimodal features for fatigue detection. The results demonstrated that there was a certain amount of complementary information between the EEG and EOG multimodality and that the information could be learned and utilized by the model. By combining the features from both modalities, we obtained a more comprehensive fatigue state indicator. Consistent with the findings of previous researchers, our multimodal fatigue detection approach yielded promising results.

Module

In previous studies, local features have often been focused on in multimodal information at the expense of global features. To solve this problem, we propose DCA and Transformer modules to extract local and global features, respectively. In order to validate the effectiveness of our proposed DCA and Transformer modules, we conducted ablation experiments on both. The results are presented in Table 3. The results outline the impact of these modules on different datasets in terms of the RMSE.

Table 3 shows that the individual applications of the ECA module and the Transformer module significantly improved the RMSE for both the SEED-VIG and SADT datasets. Specifically, for the SEED-VIG dataset, using the ECA module alone resulted in an RMSE of 0.0312, while employing the Transformer module alone further reduced the RMSE by 0.0231. Notably, the combination of both the ECA and Transformer modules yielded the best performance with the lowest RMSE of 0.0180. The situation is similar for on SADT. These results clearly indicate that while both the ECA and Transformer modules independently contribute to performance improvements, their combined usage synergistically enhances the model's accuracy, as evidenced by the lowest RMSE values obtained in both datasets.

To further demonstrate the functionality of each module, we used the uniform manifold approximation and projection (UMAP) method to visualize the process of feature extraction and the results of which are shown int Figure 3. EEG+EOG represents the original data after the input representation layer, EEG+EOG_D represents the feature after the DCA layer, and EEG+EOG_T represents the feature after the Transformer layer. Figure 3 shows that points with the same characteristics converge as the model depth increases. The results further validate the role of both the ECA and Transformer modules in feature extraction.

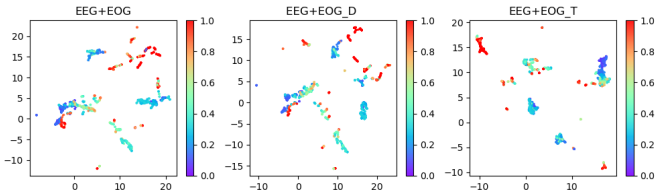

Figure 3: UMAP was used to visualize the role of feature extraction for various types of modules.

By focusing on PSD and DE features from EEG and EOG signals, the ECA has enhanced our understanding of brain and eye activities. Its adaptive channel recalibration effectively captures the dynamic nature of these signals. The ECA uses a one-dimensional convolution approach, maintaining the richness of the original features while avoiding the complexity observed in traditional attention mechanisms. The Transformer encoder outperforms the other methods, benefiting from its parallel computing capability and direct modeling of the entire input. Furthermore, the design of multi-head attention in the encoder layer further enhances the model's ability to capture different subspaces of features, thereby improving its representational power. To summarize, both modules are integral parts of the model.

Conclusion

We designed a model called DCAT that can be used for fatigue detection by effectively extracting local-global features of EEG and EOG multimodal information, and the main modules included the ECA and Transformer modules. The ECA module was utilized to extract channel-specific local features, while the transformer was employed to capture global features. The experimental results demonstrate that our model achieves state-of-the-art performance, thereby validating the complementary nature of EEG and EOG information. Furthermore, we conducted ablation experiments to further analyze the roles of each module in our model.

Acknowledgements

This work was supported by the STI 2030-Major Projects under grant 2022ZD0208900, the Guangdong Basic and Applied Basic Research Foundation under grant 2024A1515010524, and the Major Projects of Colleges and Universities in Guangdong Province under grant 2023ZDZX2021.

References

Cao, Z., Chuang, C.-H., King, J.-K., & Lin, C.-T. (2019). Multi-channel EEG recordings during a sustained-attention driving task. *Scientific data*, *6*(1), 19.

Ding, N., Zhang, C., & Eskandarian, A. (2022). EEG-Fest: Few-shot based Attention Network for Driver's Vigilance Estimation with EEG Signals. *arXiv preprint arXiv:2211.03878*.

Duan, R.-N., Zhu, J.-Y., & Lu, B.-L. (2013). Differential entropy feature for EEG-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 81–84).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Jiang, Y., Zhang, Y., Lin, C., Wu, D., & Lin, C.-T. (2020). EEG-based driver drowsiness estimation using an online multi-view and transfer TSK fuzzy system. *IEEE Transactions on Intelligent Transportation Systems*, *22*(3), 1752–1764.

Pan, J., Cai, X., Mo, D., Yu, Y., & Li, Y. (2023). Residual Attention Capsule Network for Multimodal EEG-and EOG-Based Driver Vigilance Estimation. *IEEE Transactions on Instrumentation Measurement*, *72*, 3307756.

Shi, L.-C., Jiao, Y.-Y., & Lu, B.-L. (2013). Differential entropy feature for EEG-based vigilance estimation. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6627–6630).

Song, K., Zhou, L., & Wang, H. (2021). Deep coupling recurrent auto-encoder with multi-modal EEG and EOG for vigilance estimation. *Entropy*, *23*(10), 1316.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11534–11542).

Zhang, G., & Etemad, A. (2021). Capsule attention for multimodal EEG-EOG representation learning with application to driver vigilance estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 1138–1149.

Zhang, G., & Etemad, A. (2023). Distilling EEG representations via capsules for affective computing. *Pattern Recognition Letters*, *171*, 99–105.

Zhang, Y., Guo, R., Peng, Y., Kong, W., Nie, F., & Lu, B.-L. (2022). An auto-weighting incremental random vector functional link network for eeg-based driving fatigue detection. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–14.

Zheng, W.-L., & Lu, B.-L. (2017). A multimodal approach to estimating vigilance using EEG and forehead EOG. *Journal of neural engineering*, *14*(2), 026017.