An Interpretable and Attention-Based Method for Gaze Estimation Using Electroencephalography

Nina Weng[1(✉)], Martyna Plomecka[2], Manuel Kaufmann[3], Ard Kastrati[3], Roger Wattenhofer[3], and Nicolas Langer[2]

[1] Technical University of Denmark, Kongens Lyngby, Denmark
ninwe@dtu.dk
[2] University of Zurich, Zurich, Switzerland
martyna.plomecka@uzh.ch, n.langer@psychologie.uzh.ch
[3] ETH Zurich, Zurich, Switzerland
{kamanuel,akastrati,wattenhofer}@ethz.ch

Abstract. Eye movements can reveal valuable insights into various aspects of human mental processes, physical well-being, and actions. Recently, several datasets have been made available that simultaneously record EEG activity and eye movements. This has triggered the development of various methods to predict gaze direction based on brain activity. However, most of these methods lack interpretability, which limits their technology acceptance. In this paper, we leverage a large data set of simultaneously measured Electroencephalography (EEG) and Eye tracking, proposing an interpretable model for gaze estimation from EEG data. More specifically, we present a novel attention-based deep learning framework for EEG signal analysis, which allows the network to focus on the most relevant information in the signal and discard problematic channels. Additionally, we provide a comprehensive evaluation of the presented framework, demonstrating its superiority over current methods in terms of accuracy and robustness. Finally, the study presents visualizations that explain the results of the analysis and highlights the potential of attention mechanism for improving the efficiency and effectiveness of EEG data analysis in a variety of applications.

Keywords: EEG · Interpretable model · Attention Mechanism

1 Introduction

Gaze information is a widely used behavioral measure to study attentional focus [7], cognitive control [19], memory traces [23] and decision making [28]. The most commonly used gaze estimation technique in laboratory settings is the infrared

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43895-0_69.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 734–743, 2023.
https://doi.org/10.1007/978-3-031-43895-0_69

eye tracker, which detects gaze position by emitting invisible near-infrared light and then capturing the reflection from the cornea [6]. While infrared eye tracker still remains the most accurate and reliable solution for the gaze estimation, these systems have several limitations, including individual differences in the contrast of the pupil and iris and the need for time-consuming setup and calibration before each scanning session [3,11].

Recently, Electroencephalogram (EEG) has been explored as an alternative method to estimate eye movements by recording electrical activity from the brain non-invasively with high temporal resolution [16]. The growing body of literature has shown that Deep Learning architectures could be significantly effective for many EEG-based tasks [4,26]. Nevertheless, with the advantages that Deep Learning brings, new challenges arise. Most of these models applied to electroencephalography (EEG) data tend to lack *interpretability*, making it difficult to understand the underlying reasons for their predictions, which subsequently leads to a decrease in the acceptability of advanced technology in neuroscience [25]. However, a potential solution already exists, in the form of the attention mechanism [29]. The attention mechanism has the potential to provide a more transparent and understandable way of analyzing EEG data, enabling us to comprehend the relationships between different brain signals better and make more informed decisions based on the results. With the development and implementation of these techniques, we can look forward to a future where EEG data can be utilized more effectively and efficiently in various applications.

Attention mechanisms have recently emerged as a powerful tool for processing sequential data, including time-series data in various fields such as natural language processing, speech recognition, and computer vision [5,24,29]. In the context of EEG signal analysis, attention mechanism has shown promising results in various applications, including sleep stage classification, seizure detection, and event-related potential analysis [8,13,17]. Since different electrodes record the brain activity from the different brain areas and functions, the information density from each electrode can vary for different tasks [15].

In this study, we introduce a new deep learning framework for analyzing EEG signals applying attention mechanisms. For the method evaluation, we used the EEGEyeNet dataset and benchmark [16], which includes concurrent EEG and infrared eye-tracking recordings, with eye tracking data serving as a ground truth. Our method incorporates attention modules to assign weights to individual electrodes based on their importance, allowing the network to prioritize relevant information in the signal. Specifically, we demonstrate the ability of our framework to accurately predict gaze position and saccade direction, achieving superior performance compared to previously benchmarked methods. Furthermore, we provide visualizations of model's interpretability through case studies.

2 Model

2.1 Motivation

In this study, our primary goal was to build a model sensitive to different electrodes. The motivation for this goal is two-fold. Firstly, with regards to

interpreting the model, the electrodes can be considered the smallest entity as they record signals from specific regions of the brain. Therefore, the electrode-based explanation is a reasonable approach considering human understanding. Second, in the context of model learning, incorporating adaptive weighting of electrodes within a neural network can potentially enhance the accuracy and reliability of gaze estimation systems. This is because electrodes are functionally connected to cognitive behaviors. Specifically, in tasks such as gaze estimation, electrodes positioned near the eyes can capture electrical signals from the orbicularis oculi muscles [2], thereby making the pre-frontal brain areas more crucial for precise estimation [15]. Additionally, the noise of EEG recordings could be induced by broken wire contacts, too much or dried gel, or loose electrodes [27], the influence of such electrodes should be reduced in the network under ideal circumstances.

As shown in Fig. 1, our model design focuses on enhancing an existing deep learning architecture with an electrode-sensitive component. This component first extracts electrode-related information, and then utilizes this information for two purposes: (1) emphasizing the reliable electrodes and diminishing the influence of suspicious electrodes, while simultaneously (2) providing explanations for each prediction.

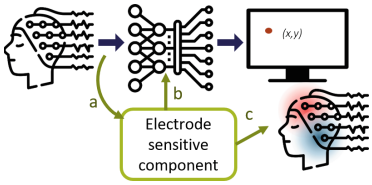


Fig. 1. We augment an electrode-sensitive component to a deep learning model, which works as follows: a) extract electrode-wise information from input data, b) control the predictions, and c) provide explanations.

2.2 Attention-CNN

Following the idea from the previous section, we propose the Attention-CNN model, where the attention blocks are used as the electrode-sensitive component. As shown in Fig. 2, the Attention-CNN model is structured by adding an attention block after each convolution block in every layer and an additional single attention block before the final prediction block (the blocks in blue). A

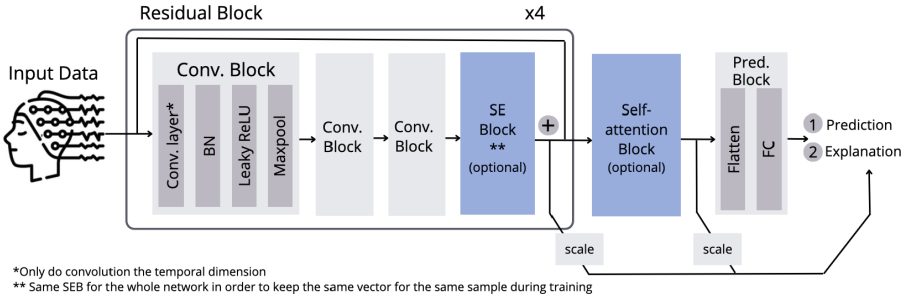


Fig. 2. The Architecture of the Attention-CNN model. (color figure online)

convolution block contains a convolution layer, a batch-norm layer [14], a leaky ReLU [18] and a max-pooling layer. In addition, the residual [10] techniques are applied in the CNN framework. The convolution layer operates only in the time dimension. The attention blocks, acting as an electrode-sensitive component, can be carried out by Squeeze-and-Excitation Block (SE Block) [12] and/or Self-Attention Block (SA Block) [29]. In the attention blocks, the retrieved electrode importance is used to weigh the features in each layer. Additionally, the same weights can provide explanations for the predictions of the model. In the prediction block, the features are flattened and then fed into the fully connected layer to finally obtain the predictions. While the SA Block is only required once in the process, the SE Blocks are added in every residual block. In order to keep the same scale for the same sample, the parameters of the SE Blocks are shared for the whole process. All building blocks are trained end-to-end, including the weights for the electrode importance used in the attention blocks.

Squeeze and Excitation Block: the SE block involves two principle operations. The **Squeeze** operation compresses features $u \in \mathbb{R}^{T' \times J}$ into electrode-wise vectors $z \in \mathbb{R}^J$ by using global average pooling. Here, T' denotes the feature size, and J is the number of electrodes. More precisely, the j-th element of z is calculated by $z_j = \mathbf{F}_{sq}(\mathbf{u}_j) = \frac{1}{T'} \sum_{i=1}^{T'} u_j(i)$. The **Excitation** operation first computes activation s by employing the gating mechanism with sigmoid activation: $s = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z}))$, where σ refers to the sigmoid function, δ represents the ReLU [20] function, and $\mathbf{W}$ are learnable weights. The final output of SE block weigh each channel adaptively by re-scaling U with s: $\tilde{\mathbf{x}_j} = \mathbf{F}_{scale}(\mathbf{u}_j, s_j) = s_j \cdot \mathbf{u}_j$. In contrast to the original implementation [12] which deals with 3-dimensional data, the input data in our setup has only 2 dimensions (electrodes and time).

Self Attention Block: The *self-attention* mechanism [22] was first used in the field of Natural language processing (NLP), aiming at catching the attention of/between different words in a sentence or paragraph. The attention is obtained by letting the input data interact with *themselves* and determining which features are more important. This was implemented by introducing the *Query, Key, Value* technique, which is defined as $\mathbf{Q} = \phi_Q(\mathbf{U}, \mathbf{W}_Q)$, $\mathbf{K} = \phi_K(\mathbf{U}, \mathbf{W}_K)$, $\mathbf{V} = \phi_V(\mathbf{U}, \mathbf{W}_V)$, where U denotes the input of self-attention block and $\phi(\cdot, \cdot)$ represents linear transformation.

Then, *Attention Weights* are computed using Query and Key:

$$\mathbf{M}_{att} = softmax(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}})$$

where d_k stands for the dimensions of the Key, and $\sqrt{d_k}$ works as a scaling factor. The softmax function was applied to adjust the range of the value in attention weights ($\mathbf{M_{att}}$) to $[0, 1]$.

Unlike the transformer model, the attention weights are first compressed into a one-dimensional vector by a layer of global average pooling (ψ) and normalized

by a sigmoid function. More precisely, we compute $\mathbf{Z}_{att} = sigmoid(\psi(\mathbf{M}_{att}))$. Finally, the output of SA Block $\mathbf{X}$ is computed by : $\mathbf{X} = \kappa(\mathbf{Z}_{att}, V)$, where κ denotes the electrode-wise production.

3 Experiments and Results

3.1 Materials and Experimental Settings

EEGEyeNet Dataset: For our experiments, we utilized the EEGEyeNet dataset [16], which includes synchronized EEG and Eye-tracking data. The EEG signals were collected using a high-density, 128-channel EEG Geodesic Hydrocel system sampled at a frequency of 500 Hz. Eye-tracking data, including eye position and pupil size, were gathered using an infrared video-based eye tracker (EyeLink 1000 Plus, SR Research), also operating at a sampling rate of 500 Hz. The recorded EEG and eye-tracking information was pre-processed, synchronized and segmented into 1-second clips based on eye movements. The infrared eye tracking recordings were used as ground truth. In this paper, the processed dataset we utilized contains two parts: the *Position Task* and *Direction Task*, which correspond to two types of eye movements: *fixation*, i.e., the maintaining of the gaze on a single location, and *saccade*, i.e. the rapid eye movements that shift the centre of gaze from one point to another. While *Position Task* estimates the absolute position from fixation, *Direction Task* estimates the relative changes during saccades, involving two sub-tasks, i.e., the prediction of amplitude and angle. The statistics and primary labels of these two parts are shown in Table 1.

Table 1. Dataset Description

Task	#Subjects	#Samples	Primary labels
Position	72	50264	`subject_id`: the identical ID of the participant `pos`: the fixation position in the form of (x, y)
Direction	72	41783	`subject_id`: the identical ID of the participant `amplitude`:the distance in pixels during the saccade `angle`: the saccade direction in radians

To ensure data integrity and prevent data leakage, the dataset was split into training, validation, and test sets across subjects, with 70 % of the subjects used for training, and 15% each for validation and testing. This procedure ensures that no data from the same subject appears in both the training and validation/testing phases, thereby avoiding potential subject-related patterns from being learned by the model during training and tested on in validation/testing. For more details of this dataset, please refer to [16].

Implementation Details: The experiments are implemented with PyTorch [21]. When training the Attention-CNN model, the batch size is set to 32, the number of epochs is 50, and the learning rate is $1e^{-4}$. There are 12 convolution

blocks, and the residual operation repeats every three convolution blocks. The feature length of the hidden layer is set as 64, and the kernel size is 64. The number of convolutional layers, kernel size and hidden feature length, are selected based on validation performance. We conducted experiments with three configurations: the SE Block and the SA Block together, only one of the attention blocks, or no attention blocks at all. For the angle prediction in Direction Task, we use angle loss $l_{angle} = |(atan(sin(p-t), cos(p-t))|$, where p denotes the predicted results, and t denotes the targets. For Position Task and Amplitude prediction in the Direction Task, the loss function is set to smooth-L1 [9].

Evaluation: For Position task, Euclidean distance is applied as the evaluation metric in both pixels and visual angles. Compared to pixel distance, visual angles depend on both object size on the screen and the viewing distance, thus enabling the comparison across varied settings. The performance of Direction Task is measured by the square root of the mean squared error (RMSE) for the angle (in radians) and the amplitude (in pixels) of saccades. In order to avoid the error caused by the repeatedness of angles in the plane (i.e. 2π and 0 rad represents the same direction), $atan(sin(\alpha), cos(\alpha))$ is applied, just like in angle loss.

3.2 Performance of the Attention-CNN

Table 2 shows the quantitative performance of the Attention-CNN in this work. For the Position Task, CNN with SE block has an average performance with the RMSE of 109.58 pixels. Likewise, the CNN model with both SE block and the SA block has a similar performance (110.05 pixels). Similar to Position Task, in amplitude prediction of Direction Task, the attention blocks aid the prediction evidently, heightening the performance by 5 pixels. Here, the model with both attention blocks has a lower variance. For angle prediction, the CNN model with both SE block and SA block has the best performance among all with the RMSE of 0.1707 rad.

We can conclude that the CNN model with both attention blocks consistently outperforms the CNN model alone by 5 to 10 percent across all tasks, indicating that electrode-wise attention assists in the learning process of the models.

Table 2. The performance of the Attention-CNN on Direction and Position Task.

Models	*Angle/Amplitude*		*Abs. Position*
	Angle RMSE	**Amp. RMSE**	**Euclidean Distance (Visual Angle)**
CNN	0.1947 ± 0.021	57.4486 ± 2.053	115.0143 ± 0.648 (2.39 ± 0.010)
CNN + SE	0.1754 ± 0.007	55.1656 ± 3.513	$\mathbf{109.5816 \pm 0.238}$ $\mathbf{(2.27 \pm 0.004)}$
CNN + SA	0.1786 ± 0.010	$\mathbf{52.1583 \pm 1.943}$	112.3823 ± 0.851 (2.33 ± 0.013)
CNN + both	$\mathbf{0.1707 \pm 0.011}$	$\mathbf{52.2782 \pm 1.169}$	$\mathbf{110.0523 \pm 0.670}$ $\mathbf{(2.28 \pm 0.010)}$

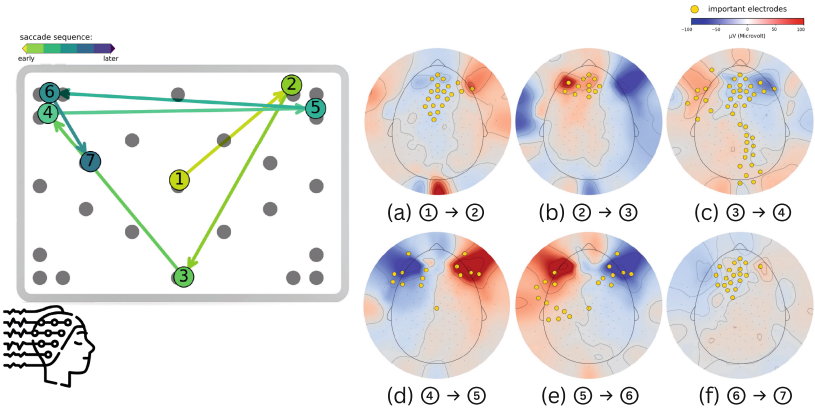


Fig. 3. Visualization of signal intensity across scalp and electrode importance from our models. Left: the track of a continuous sequence of saccades. Right: the corresponding brain activities (red: positive electrical signal, blue: negative electrical signal) and the important electrodes detected by the attention-based model (denoted as yellow nodes, the threshold is set as the mean value of all electrodes during the sequence). The model used here is the CNN with SA block. (Color figure online)

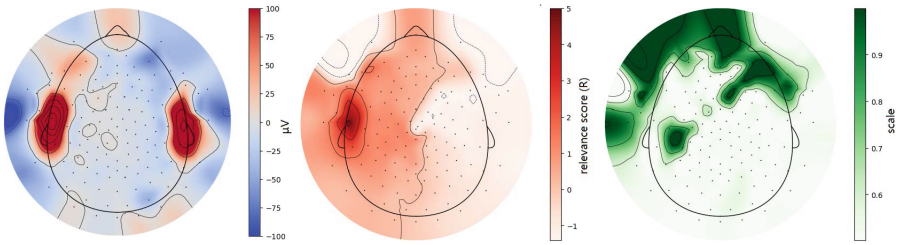
3.3 Model Interpretability by Case Studies

To provide a more detailed analysis of the interpretability of our proposed Attention-CNN model, as well as to further investigate the underlying reasons for the observed accuracy improvement, we conducted a visual analysis of the model performance, with a particular focus on the role of the attention block. Our analysis yielded two key findings, which are as follows:

Firstly, the attention blocks were able to detect the electrical difference between the right and left pre-frontal area in case of longer saccades, i.e. rapid eye movements from one side of the screen to the other; see the saccades (d) and (e) in Fig. 3. We present the sequence of saccades and observed the EEG signals as well as the electrode importance from proposed models in Fig. 3. The attention block effectively captured this phenomenon by highlighting the electrodes surrounding the prominent signals (saccades (d) and (e) in Fig. 3). Conversely, in cases where the saccade was of a shorter distance (other saccades in Fig. 3), attention was more widely distributed across the scalp rather than being concentrated in specific regions. This is justifiable as the neural network aims to integrate a more comprehensive set of information from all EEG channels.

Additionally, the attention block effectively learned to circumvent the interference caused by noisy electrodes and redirected attention towards the frontal region. Figure 4 illustrates a scenario where problematic electrodes were situated around both ears, exhibiting abnormal amplitudes ($\pm 100\ \mu V$). Using Layer-wise Relevance Propagation [1] to elucidate the CNN model's predictions, the result depicted in Fig. 4b revealed that the most significant electrodes were located over the left ear, coinciding with the noisy electrodes. In contrast, as shown in

Fig. 4c, the Attention-CNN model effectively excluded the unreliable electrodes and allocated greater attention to the frontal region of the brain.



(a) Input EEG data (b) LRP results from CNN (c) Scales from CNN+SA

Fig. 4. One example of test samples containing problematic electrodes is the Position Task. As shown in (a), the dark red areas around the ears represent intense electrical signals with abnormal amplitudes (>100 V). In (b), the Layer-wise Relevance Propagation (LRP) results from the CNN model reveal that the electrodes around the left ear still play a crucial role in the prediction process. Conversely, the Attention-CNN model's results (c), indicate that it bypasses the ear area and allocates more emphasis to the pre-frontal region. As a result, the error in Euclidean Distance improved by 200.85 pixels for this specific sample (from 265.18 to 64.33). (Color figure online)

3.4 Explainability Quantification

We further examine the validity in explainability of the proposed method by comparing the distribution of learned attention of noisy and non-noisy electrodes in the Direction Task. The attention block's effectiveness is demonstrated by its ability to assign lower weights to these noisy electrodes in contrast to the non-noisy ones. Within all samples in the Direction Task that feature at least one noisy electrode, only 19% of the non-noisy electrodes had normalized attention weights below 0.05. In contrast, 42% of the noisy electrodes exhibited this trait, implying the attention block's ability to reduce weights of abnormal electrodes. We direct readers to the Supplementary materials for a distribution plot showcasing the difference between noisy and non-noisy electrodes, along with additional details. It's important to note that quantifying explainability methods for signal-format data, such as EEG, presents a significant challenge and has limited existing research. Therefore, additional investigations in this field are anticipated in future studies.

4 Conclusion

In this study, we aimed to address the issue of the lack of interpretability in deep learning models for EEG-based tasks. Our approach was to leverage the

fact that EEG signal noise or artifacts are often localized to specific electrodes. We accomplished this by incorporating attention modules as electrode-sensitive components within a neural network architecture. These attention blocks were used to emphasize the importance of specific electrodes, resulting in more accurate predictions and improved interpretability through the use of scaling.

Moreover, our proposed approach was less susceptible to noise. We conducted comprehensive experiments to evaluate the performance of our proposed Attention-CNN model. Our results demonstrate that this model can accurately classify EEG and eye-tracking data while also providing insights into the quality of the recorded EEG signals. This contribution is significant as it can lead to the development of new decoding techniques that are less sensitive to noise.

In summary, our study underscores the importance of incorporating attention mechanisms into deep learning models for analyzing EEG and eye-tracking data. This approach opens up new avenues for future research in this area and has the potential to provide valuable insights into the neural basis of cognitive processes.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
2. Bulling, A., Ward, J.A., Gellersen, H., Tröster, G.: Eye movement analysis for activity recognition using electrooculography. IEEE Trans. Pattern Anal. Mach. Intell. **33**(4), 741–753 (2010)
3. Carter, B.T., Luke, S.G.: Best practices in eye tracking research. Int. J. Psychophysiol. **155**, 49–62 (2020)
4. Craik, A., He, Y., Contreras-Vidal, J.L.: Deep learning for electroencephalogram (EEG) classification tasks: a review. J. Neural Eng. **16**(3), 031001 (2019)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Duchowski, A., Duchowski, A.: Eye tracking techniques. eye tracking methodology: Theory Pract. 51–59 (2007)
7. Eckstein, M.K., Guerra-Carrillo, B., Singley, A.T.M., Bunge, S.A.: Beyond eye gaze: what else can eyetracking reveal about cognition and cognitive development? Dev. Cogn. Neurosci. **25**, 69–91 (2017)
8. Feng, L.X., et al.: Automatic sleep staging algorithm based on time attention mechanism. Front. Hum. Neurosci. **15**, 692054 (2021)
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Holmqvist, K., Nyström, M., Mulvey, F.: Eye tracker data quality: what it is and how to measure it. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 45–52 (2012)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

13. Hu, Z., Chen, L., Luo, Y., Zhou, J.: EEG-based emotion recognition using convolutional recurrent neural network with multi-head self-attention. Appl. Sci. **12**(21), 11255 (2022)
14. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456. PMLR (2015)
15. Kastrati, A., Plomecka, M.B., Küchler, J., Langer, N., Wattenhofer, R.: Electrode clustering and bandpass analysis of eeg data for gaze estimation. arXiv preprint arXiv:2302.12710 (2023)
16. Kastrati, A., et al.: Eegeyenet: a simultaneous electroencephalography and eye-tracking dataset and benchmark for eye movement prediction. arXiv preprint arXiv:2111.05100 (2021)
17. Lee, Y.E., Lee, S.H.: EEG-transformer: Self-attention from transformer architecture for decoding eeg of imagined speech. In: 2022 10th International Winter Conference on Brain-Computer Interface (BCI), pp. 1–4. IEEE (2022)
18. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of ICML. vol. 30, p. 3. Atlanta, Georgia, USA (2013)
19. Munoz, D.P., Everling, S.: Look away: the anti-saccade task and the voluntary control of eye movement. Nat. Rev. Neurosci. **5**(3), 218–228 (2004)
20. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
21. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. Adv. Neural Inform. Process. Syst. **32** (2019)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
23. Ryan, J.D., Riggs, L., McQuiggan, D.A.: Eye movement monitoring of memory. JoVE (J. Visualized Exp.) (**42**), e2108 (2010)
24. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
25. Sturm, I., Lapuschkin, S., Samek, W., Müller, K.R.: Interpretable deep neural networks for single-trial EEG classification. J. Neurosci. Methods **274**, 141–145 (2016)
26. Tabar, Y.R., Halici, U.: A novel deep learning approach for classification of EEG motor imagery signals. J. Neural Eng. **14**(1), 016003 (2016)
27. Teplan, M., et al.: Fundamentals of EEG measurement. Measure. Scie. Rev. **2**(2), 1–11 (2002)
28. Vachon, F., Tremblay, S.: What eye tracking can reveal about dynamic decision-making. Adv. Cogn. Eng. Neuroergonom. **11**, 157–165 (2014)
29. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)