Spatio-Temporal Analysis of Transformer based Architecture for Attention Estimation from EEG

Victor Delvigne * †, Hazem Wannous †, Jean-Philippe Vandeborre †, Laurence Ris ‡, Thierry Dutoit *
* ISIA Lab, Faculty of Engineering, *University of Mons*, Belgium;
† IMT Nord Europe, *CRIStAL UMR CNRS 9189*, France;
‡ Neuroscience Lab, Faculty of Medecine and Pharmacy, *University of Mons*, Belgium
Email: victor.delvigne@umons.ac.be

***Abstract*—For many years now, understanding the brain mechanism has been a great research subject in many different fields. Brain signal processing and especially electroencephalogram (EEG) has recently known a growing interest both in academia and industry. One of the main examples is the increasing number of Brain-Computer Interfaces (BCI) aiming to link brains and computers. In this paper, we present a novel framework allowing us to retrieve the attention state, i.e degree of attention given to a specific task, from EEG signals. While previous methods often consider the spatial relationship in EEG through electrodes and process them in recurrent or convolutional based architecture, we propose here to also exploit the spatial and temporal information with a transformer-based network that has already shown its supremacy in many machine-learning (ML) related studies, e.g. machine translation. In addition to this novel architecture, an extensive study on the feature extraction methods, frequential bands and temporal windows length has also been carried out. The proposed network has been trained and validated on two public datasets and achieves higher results compared to state-of-the-art models. As well as proposing better results, the framework could be used in real applications, e.g. Attention Deficit Hyperactivity Disorder (ADHD) symptoms or vigilance during a driving assessment.**

I. Introduction

Nowadays, deep learning (DL) and other ML algorithms have known a huge increase in interest that has led to improvements in several scientific fields. Different domains have benefited from ML researches such as natural language processing (NLP), computer vision, speech recognition or understanding. However, another field where the use of DL remains elusive is brain imaging, the goal of these models being to help to better understand the mechanism within the brain. The considered signals can represent the brain matter composition with magnetic resonance imaging (MRI) [1] or the electrical activity of its neurons with EEG [2]. The goal of the model considering EEG are to evaluate human cognitive faculties to have a better understanding of brain function.

Recent works propose to consider electrophysiological signals and especially EEG to estimate the attentional state of participant [3]–[6], i.e. a metric expressing the ability of an individual to be concentrate on a given task. The purposes of these researches are wide and could help in various fields: medical, entertainment, road safety or marketing. The proposed approaches consider the use of ML models that have already shown outperforming results in other fields: fully-connected neural networks [7], convolutional neural networks (CNN) [6], recurrent neural networks (RNN) [3] or combinations of these last. Moreover, the models can consider extracted features from EEG [5] or preprocessed EEG directly [6].

Although these works present promising results, they tend to ignore some of the sequential relationships governing EEG signals. This sequential relationship being modelled in the temporal (EEG signals can be considered as a set of time series), spatial (EEG are recorded in several locations of the participant scalp) and frequential (EEG can be filtered in different frequency bands each of them being responsible for human behaviour).

In this context, due to the encouraging results of the novel techniques aiming to improve the analysis of sequential information: written sentences [8] or speech segments [9], it has been thought to merge this advanced to improve participant attention estimation from EEG. Our approach is based on the self-attention transformer encoder layers [10] allowing us to combine information from a non-neighbour element in a sequential signal which was not the case with conventional RNN as reported in [11]. Transformer based model can automatically process the sequential information from frequential bands, temporal windows or electrode location. In this work, feature matrices have been extracted from EEG data and represented in a 3D frame with three specific dimensions: temporal, spatial and frequential. This novel matrices representation is specially dedicated to our transformer architecture and aims to estimate attention state. The contributions of this work are the following: 1) creating transformer-inspired architecture suiting with EEG; 2) developing a three-stream network aiming to estimate attention state; 3) assessing the effect of frequential bands, temporal windows length and electrodes location; 4) finally, proposing a method that presents encouraging results exceeding the state-of-the-art approaches; 5) making extensive analysis of the feature extraction methods.

II. Related Work

During the last decade several research projects considering EEG signals have been completed. These last use ML algorithms for different estimation [7], [12]–[17]. One of the specific subset of models consider the use EEG to retrieve the attentional or vigilance state of participants, i.e. focus vs. distract [7], [18], [19].

Commonly, the first step consists of the preprocessing corresponding to band-pass filtering with or without ocular artefacts removing methodology. After, feature extraction often consists to frequential feature extraction [5], [12], [15], [20]. However, some research projects proposed also an approach based on an automatic feature extraction methodology made with other DL models, i.e. deep-autoencoder [21], or with larger DL architecture to automatically extract features for classification/regression [14]. Although this method is mainly employed in many other fields, it remains difficult to consider EEG processing without a handcrafted feature extraction step due to the signals' nature and the relatively small size of the public datasets. Most of the proposed approaches consider frequential based feature extraction methods to process EEG, however, other methods reflecting different signal properties can also be considered: signal's disorder with fractal dimension [22], temporal domain properties [23], [24].

From the computed feature arrays, it exists many representations based on methods originally dedicated to other tasks that have been adapted to EEG, e.g. image [15], [25] or graph [13]. The proposed representation in this paper will take into account the interdependence that exists among EEG signals in the spectral, temporal and spatial domains.

One of the main challenges this paper aims to tackle is the management of the sequential aspect of the considered inputs, i.e. considering the best ML-based approach to benefit from the relationship among input signals. In particular, this challenge is to find the best way to express the spatial (through electrodes), frequential (through frequential bands) and temporal (through the signal's temporal evolution) relationship between EEG signals. Among the existing work, different methodologies have been proposed to solve these issues but often consider only the spatial information, i.e. how to organise the information to consider electrodes positions on the scalp:

- Recurrent Neural Networks (incl. LSTM and GRU) that process spatial information in a unidirectional pathway. It is then necessary to consider one RNN for each direction (EEG spatial relationship being in two dimensions). Moreover, these models aim to estimate the recurrence in the sequential information, in the case of longer sequences the relationship between elements too far apart may not be taken into account [11].
- Convolutional Neural Network can be used to model the spatial [26], [27] by considering a 2D representation of EEG feature matrices [26]. An improved method aims to take into account the position of the electrodes by creating an image based on the interpolation of the location of the electrodes in the 3D frame [15], [25]. Another approach consists to extract temporal information from raw signals in the temporal and spatial domain by considering two-dimensions kernels [14].
- Graph neural networks are a type of neural network that considers inputs as a graph. In the context of EEG, each electrode is considered as a node and the edges are proportional to the distances between them [13].

It is important to note that the above-mentioned approaches are not necessarily implemented straightforwardly. Some works proposed a different approach consisting of concatenation or parallelized models. On the other hand, it exists methodologies considering a novel approach to improve the baseline results. For instance, by considering images-based EEG with CNN but with the help of self-attention mechanism on the spatial and temporal stream to increase the classification accuracy [26].

Although several already presented approaches show high accuracy for EEG classification/regression in most of the cases they only consider an interpolated or uni-dimensional relationship between sequential information. For this reason, it has been thought to consider a novel approach for attention estimation by considering the sequential aspects in three directions: spatial, temporal and frequential.

III. Proposed Method

In this paper, we proposed an innovative model aiming to estimate the attention state from EEG. The proposed approach to estimate attention is inspired by the transformer encoder from self-attention based model [10]. The motivations behind the use of this kind of model are justified by their ability to extract sequential information from different modalities [8]–[10]. The proposed pipeline is separated into four steps: signals preprocessing, segmentation and representation; features extraction; modalities classification.

The preprocessing step follows the general recommendations for reproducible EEG research [28]. The EEG dataset can be considered as a set of segmented signals $X^r = [S_1^r, S_2^r, \ldots, S_C^r] \in \mathbb{R}^{C \times T}$ with C and T representing respectively the amount of electrodes on EEG recorder and the length of the signal. A bandpass filtering has been applied on each segment between 0.5 and 50 Hz. The lower band removes the continuous contribution and detrends the signals, the higher band removing electrical artefacts oscillating at 50 Hz and a part of the muscular artefacts. An FIR filter with a Hanning window of 1-second has been considered for bandpass filtering. Another removing artefact methodology consisting of a manual removing signals by visual inspection and the use of the Automatic Artefact Removal (AAR) plug-in from EEGLab [29] has been applied to remove the remaining ocular and muscular artefacts in X_r. This step is repeated for each trial and electrode, the preprocessed dataset can be reformulated as a matrix of dimension $[n_{trials} \times C \times T]$ with n_{trials} being the number of trials during the total acquisition.

On the other hand, to compare signals corresponding to a high/low attention state, it is necessary to compute a label representing this feature. For this purpose, one physiological measurement correlated with the attention state have been considered: the reaction time during sustained-attention task, i.e. the time taken by a participant to react to a stimulus has been recorded. We consider the median for all the trials participants dependent (trials corresponding to the participant) and independent (all the trials). Then, a threshold is deduced for each participant by computing the mean between the

median participant dependent and independent. Finally, in a trial corresponding to a physiological measurement above (resp. below), the threshold is considered as a low (resp. high) attention state. A binary class is then assigned to each trial.

As spectral information plays an important role in attention estimation [3], [5], it has been though to filter the signals into several frequency intervals, the latter may be physiologically pre-defined frequency bands (i.e. $\delta, \theta, \ldots$ bands) or regular spectrum decomposition between 0 and the cut-off frequency. Finally, the preprocessed EEG dataset is re-expressed by considering the band filtering as a set of signal $X_f^r = [S_{1,f}^r, S_{2,f}^r, \ldots, S_{C,f}^r] \in \mathbb{R}^{F \times C \times T}$ with $S_{i,j}^r$ being the EEG segment of i-th channel and j-th frequency band and F being the amount of considered frequential bands. The filtering being made with same filter parameters.

After separating EEG into frequential contributions, EEG segments have been segmented into time windows. The goal of this segmentation is to capture the information from the signal's temporal evolution during task processing. Studies have shown that specific patterns occur in EEG during the sight of a stimulus [14]. The novel signal representation is $X_f^t = [S_{1,f}^t, S_{2,f}^t, \ldots, S_{C,f}^t] \in \mathbb{R}^{F \times C \times T' \times \frac{T}{T'}}$ with T' being the amount of temporal window, the length of the segment after the temporal segmentation is $\frac{T}{T'}$.

Then, from X_f^t, feature are extracted to express the signal in a shorter subspace. In the context of attention estimation, different feature extraction methods have already been considered each of them considering specific signals' aspect: Differential Entropy (DE) [20]; Fisher Information (FI) [30]; Hjorth parameters [23]; Petrosian Fractal Dimension [31]; Teager Energy [32].

Finally, from the feature matrice $F \in \mathbb{R}^{F \times C \times T' \times n_{feat}}$, it is possible to consider its representation as a sequence in three dimensions: frequential, temporal and spatial. The frequential direction takes into account the signal evolution among the considered frequential bands. The spatial direction represents the electrodes based relationship between features information and depends on the order in which the electrodes are sorted. Finally, the temporal dimension expresses the time-based evolution of each feature vector. Practically, this representation can be expressed considering each of this dimension, after transposing and merging axes, the resulting representation can be expressed as: $EEG|_{frequency} \in \mathbb{R}^{F \times n_{feat-freq}}$, $EEG|_{temporal} \in \mathbb{R}^{T' \times n_{feat-temp}}$, $EEG|_{spatial} \in \mathbb{R}^{C \times n_{feat-spat}}$.

The feature dimension for each of these three representations is deduced from the reshaped dimension of the feature matrice F. The interest of this representation is that it permits to have a sequential representation of information considering each stream (i.e. frequential, temporal and spatial) separately. Moreover, this representation allows not to lose information or limit biases, unlike for instance the image-based representation that considers an interpolation of a feature map.

It exists several DL based algorithms to estimate modalities from this particular representation of information [33]. In this paper, an adapted version of the encoder layers from the transformer architecture [10] is proposed. This architecture is composed of different blocs each of them being responsible for a specific aspect. If we consider the input feature matrix F representing the sequential information in one of the three dimensions as explained in the previous subsection, the estimated class $\hat{y}$ is computed after the following steps:

- Embedding aiming to have a continuous representation of the feature in a vector of lower dimension;
- Positional encoding allowing to add information about the element position in the sequence given that self-attention mechanism providing not information about recurrence in the sequence (unless RNN).
- Transformer encoder applying attention mechanism on all the previous elements composing the sequence resulting in a hidden representation of the input embedded vector.
- Adapted transformer decoder consisting of a feed-forward network (FFN) applied on the hidden vectors resulting from the encoder. A second FFN merges these vectors to provide a single representation for each sub-model corresponding to signal representation as seen in Fig 1.
- Finally, the resulting three vectors are concatenated and passed through a FFN aiming to estimate attention state.

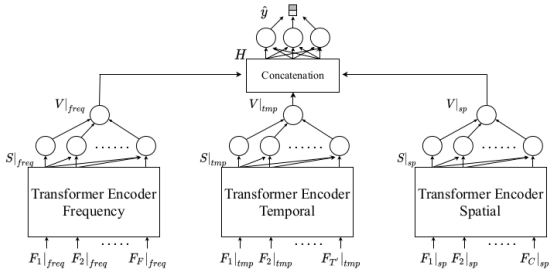


Fig. 1: Overview of the proposed architecture for attention estimation. The three representations of the EEG features arrays are passed to the correspond transformer encoder stream. Then the outputs are decoded and concatenated to create an hidden vectors passed to a fully-connected networks to estimate the attention state.

The proposed architecture aims to estimate attention state from feature arrays computed from EEG. The goal of the training phase is to find the correct value for each trainable parameter to make the correct estimation. The considered loss consists of a categorical cross-entropy.

IV. Experiments

In this section, we describe the considered datasets and models in this paper, as well as the settings and parameters for attention estimation from EEG.

A. Datasets

In our experiments, two different datasets of EEG signals have been considered: PhyDAA [3] and Driving EEG [4]. Their goal is to proposed segmented signals corresponding to a specific attention state. The methodology employed to assess attention state is based on the registration of the reaction

Approach	Driving EEG [4] ACC/STD [%]	PhyDAA [3] ACC/STD [%]
TCA + LR [18]	72.70/9.42	-
MIDA [19]	73.01/9.17	-
Graph Network [3]	-	72.41/5.51
SVM*	68.09/9.55	64.61/9.22
RF*	67.81/10.17	61.55/9.79
RNN*	72.12/8.27	70.86/9.82
ResNet*	62.07/6.20	66.82/5.21
Transformer*	**74.41/9.27**	**77.24/6.11**

TABLE I: Classification performance of the different methods considering participant independent cross-validation, i.e. with leave one subject out cross-validation accuracy. * denotes the results obtained from our models experiments.

time to specific stimuli. These stimuli can be represented by a balloon appearing inside of virtual reality (VR) environments [3] or by steering angle modification of a car during a driving task [4]. PhyDAA dataset proposed an experiment during which participants are asked to react as fast as possible to a specific stimulus. The reaction is measured with the direction of the eyes, it corresponds to the time elapsed to direct the sight toward the stimuli. 32 participants took part in the 15 minutes length experiments. Driving EEG dataset consists of an attention assessment during driving task [4]. This dataset proposes a task in VR environment representing a car driving task, during which it is asked to react as fast as possible to perturbators corresponding to the deviation of the car trajectory. The time taken to correct the steering angle is jointly measured. 27 participants took part in the 90 minutes experiment.

The 32 electrodes have been placed following the 10/20 disposition for both datasets and registered at a sampling frequency of 500 Hz. The steps already presented in the third section have been applied to extract the feature and split the feature arrays in each of the three directions (frequency, temporal and spatial). To investigate the effect of frequential bands and temporal windows, it has been decided to divide the samples into $[1, 4, 10, 20]$ temporal windows and have been filtered in $[1, 5, 20, 50]$ frequential bands[1].

B. Settings

The model evaluates with subject dependent and independent has been configured with the same parameters. The chosen dimensions were respectively equal to $F = [1, 5, 20, 50]$, $T' = [1, 4, 10, 20]$ and $C = 32$ for each of the three dimensions. The transformer encoder part of the architecture is composed of two transformer encoder layers each of them composed of four heads in the multi-attention model part [10]. The chosen dimension for the embedded representation and the dimension of the self-attention matrices is equal to 64 and 128. The training has been made considering a stochastic gradient descent (SGD) optimizer with a scheduled learning rate beginning at $1e-2$ with $\gamma = 0.99$. The batch size and number of epochs are respectively 32 and 250. The model has been implemented using Pytorch library and the training has been made on one Nvidia Titan RTX GPU. For sake of reproducibility, the model's implementation and the codes used for the preprocessing are freely available on github[2].

[1]The computed value for the lengths of both windows have been chosen after a preliminary study.

V. Results

In this section, we discuss the results achieved to retrieve the attention state. A comparison with other methodology, a study of the different chosen parameters, the activation maps resulting and an ablation study has been performed.

A. Comparison of deep learning models

To evaluate the proposed methodology, the architecture has been trained and validated with two different datasets. Two training methodologies aiming to assess the model faculty to generalise have been considered in this paper: 1) Subject-Independent classification, where the model is trained with all the participant signals except one that is used for the validation and the step is repeated for all subjects and a mean cross-validation accuracy and its standard deviation is computed. The benefit of this method is to measure the model ability to generalise its knowledge to never met participants; 2) Subject-dependent classification where the model is trained and validated with the same participant following a regular 5-fold cross-validation, the process is repeated for each participant and the mean and standard deviation of cross-validation accuracy are computed. The advantages of this method are that it gives a good insight into the model ability to make estimations with fewer signals.

It was also thought to consider the comparison with the different methodology aiming to estimate attention from feature matrices constructed from EEG signals. Among the existing ML models, four have been considered:

- Traditional ML models: Random Forest (RF) and Support Vector Machine (SVM) based classifier to define a baseline result for attention estimation.
- RNN based approach consists of the transformer-based approach represented in Figure 1 where the transformer encoder layers are replaced by RNN for each stream.

[2]https://github.com/VDelv/Spatio-Temporal-EEG-Analysis

Approach	Driving EEG [4] ACC/STD [%]	PhyDAA [3] ACC/STD [%]
MLP [7]	81.32/6.02	-
Graph Network [3]	-	77.34/10.24
SVM*	76.07/9.65	70.82/13.25
RF*	75.60/8.76	75.63/12.89
RNN*	80.03/8.09	79.64/10.55
ResNet*	75.96/8.98	70.39/6.91
Transformer*	**83.31/6.71**	**85.04/7.56**

TABLE II: Classification performance of the different methods considering participant dependent cross-validation. * denotes the results obtained from our models experiments.

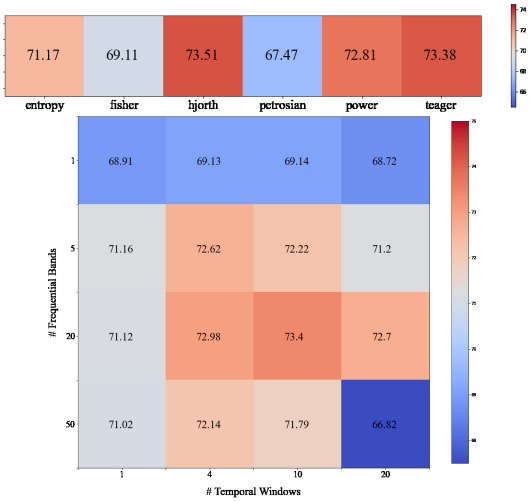


Fig. 2: Mean cross-validation accuracy in function of the feature extraction methods (above); the amount of temporal windows and frequential bands (below).

- CNN approach based on an image-based representation of the EEG feature map. Then the images are passed through a resnet architecture [34].

As seen in Tables I and II, results acquired from the transformer-based approach present the highest accuracy compared to other baseline approaches for both datasets that demonstrate the proposed framework's ability to estimate attention from EEG. More, it shows the efficiency of self-attention based models architecture to process sequential signals.

Furthermore, results from previous works have also been compared to evaluate the proposed methodology. For the first dataset, the specificity of the previous approach is based on a transfer learning approach to increase cross-subject accuracy. These last are based on Transfer Component Analysis (TCA) or Maximum Independence Domain Adaptation (MIDA) with traditional ML architecture that may cause an accuracy decay compared to the more complex methods. As seen in Tables I and II the results from our experiments from traditional ML approaches, i.e. SVM and RF, are lower compared to other DL methods. It makes us think that considering a more complex training methodology, including transfer learning, may lead to an increase in the transformer accuracy, although its accuracy is outperforming the state-of-the-art models.

For the second dataset, the best results from the related works are based on Graph Convolution Network (GCN). Its architecture is composed of graph convolution and a pooling operation aiming to keep only the most discriminant electrodes. Unlike the transformer, GCN only considers the spatial stream to estimate attention from EEG. This may explain the lower results.

B. Feature parametetrs analysis

As mentioned in Section III, different feature extraction methods and segmentation parameters have been considered. In this section, we present the corresponding cross-validation accuracy for each combination. In Figure 2, the feature extraction methods present cross-validation accuracy around similar range of value $\approx 70\%$. As shown, Hjorth, TE, PSD and DE present the best results. Moreover, the two best feature extraction methods: Hjorth parameters and TE based operator, consider both the signals' derivative that corroborates the fact that the derivative play an important role in the attention estimation from EEG.

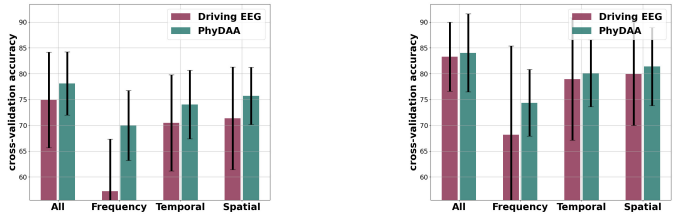
As seen, the amount of both time windows and frequency bands play an important role in attention estimation. As seen in Figure 2, for both number of temporal windows and frequential bands a too small number of time windows/frequency bands leads to a decrease of accuracy. This decay can be caused by the difficulty of representing the evolution of the brain activity during the stimuli apparition or among the spectrum. More, a too large number can lead to a decrease in accuracy due to overfitting issues.

The medium values present the higher results temporal and spectral parameters. More precisely, better results are proposed for regularly cut bands (i.e. with 20 for #Frequency Bands) compared to pre-defined bands (i.e. with 5 for #Frequency Bands). This can be explained by the fact that some populations do not present the same bands limits the pre-defined [35], [36].

C. Ablation studies

In addition to the comparison based on the considered architecture or signal parameters, it has been thought to consider a comparison aiming to investigate the contribution of each stream. For this purpose four different architectures have been considered: 1). the original approach as described in section III considering the concatenation of the three streams; 2). frequential; 3) temporal; 4) spatial based transformer stream standalone. The experimental results from these four different approaches are listed in Figure 3. As seen, in both of the cases and datasets, the best results were noted for the approach considering the information from the three directions. This observation corroborates the fact that considering all the available information is the best approach to estimate attention from EEG.

The best results for a single direction based approach were acquired by considering only the spatial stream. The temporal based approach presented slightly lower results. These findings may be explained by the fact that the spatial, i.e. electrodes-based, and temporal information have played an important role in the attention estimation. The activation areas deduced from EEG are considered as a good biomarker for behaviour/movements estimation [37], which may explain the high accuracy provided by spatial information. The importance of the temporal information and the resulting scores are explained by the nature of the signals composing the datasets. In both of them, each segment can be considered as Event-related potentials (ERP), i.e. the brain response resulting from a stimulus [38] and present a specific pattern. In the context of this experiment, stimuli apparition is fixed at $t = 1$ second,



(a) Subject Independent. (b) Subject Dependent.

Fig. 3: Ablation study for the transformer models. The left (resp. right) figure correspond to the subject independent (resp. dependent) mean cross-validation. The bars colour correspond to the considered datasets with each bar corresponding to a stream.

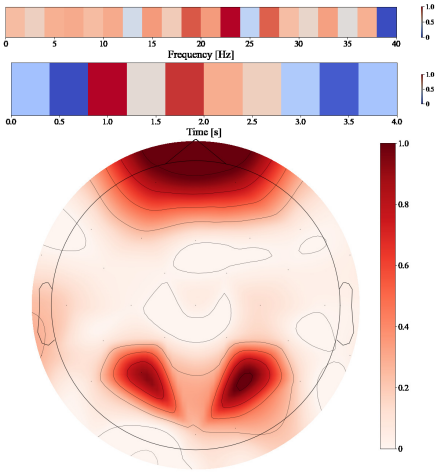


Fig. 4: Activation map in function of attention state classification of the hidden representation of the corresponding representation: Top - Spectral; Middle - Temporal; Bottom - Spatial. These maps have been normalized to reflect the contribution of each temporal windows.

however, ERP may appear at different instants depending on the attention state. The frequential based approach acquired the lowest accuracy for both dataset and training methods as mentioned in Figure 3. The consideration of this stream is motivated by the fact that previous works mentioned the importance of frequential information to retrieve attention state, and especially the middle bands [39]. Poorer results can be expressed by the redundancy of the spectral information already extracted during feature extraction.

D. Activity maps analysis

To investigate the contribution of each stream, it has been thought to analyze the activity maps generated by the transformer-encoder. For this purpose we consider the L_2 norm of each output sequence from transformer encoder for each stream, i.e. $S_i|_{freq}, S_i|_{temp}, S_i|_{spatial}$ as shown in Figure 1. This process has been made during the training of the three-stream model training and the weight have been frozen. After considering the norm of every sequence, they have been normalized to study their contribution.

In Figure 4, the activity map resulting for the spectral stream is displayed. As seen, three frequential areas emerge from the activity spectrum. First, the sparsity of the high-frequency contribution (i.e. > 25 Hz) and their irregularity among the area makes the author think that they are caused by the remaining artefacts from muscular activity that have not been removed from the pre-processing step but can be correlated with attention state as it has already been proven [40]. The two other spectral bands with high activity superposed with the physiologically defined frequency bands θ, α and β that are often used for attention estimation [18], [19], [35]. More, the behaviour related to the task proceeds by the participant is related to the theoretically defined behaviour by this frequency bands [41] that corroborates these insights.

As seen in Figure 4, the most salient instant during the trial is between 1 and 2.75 seconds with higher importance between 1 and 2 second that corresponds to the instant directly following the stimulus apparition. This period seems therefore important to distinguish high/low attention segments.

At the bottom of Figure 4, the most salient EEG-based region are displayed. First, a casi symmetry is observed between the two hemispheres that make it possible to reject an electrode misplacement or mis-conduction due to the specific registration conditions. More significantly, two different regions stand out from the spatial activity: frontal and parietal regions from electrodes placements. In addition to being one of the most salient electrodes regions for attention estimation, it has been shown that the parietal region is also responsible for attention mechanism [42].

VI. Conclusion

In this work, we present a framework aiming to estimate the attention state from EEG signals during specific tasks. We propose a novel approach to handle these signals based on a three-fold information representation based on the frequential, temporal and spatial features. Moreover, a novel Transformer inspired architecture for EEG processing has been presented. This last allows extracting the sequential information from the EEG feature maps in each of the three dimensions mentioned above. To validate this new method, the framework has been trained and tested on public datasets assessing the attention state. The results are encouraging and outperform the state of the art approaches. The proposed models can be useful for different applications such as attention assessment for subjects with ADHD to detect and help to reduce their symptoms; another application could be a vigilance estimator during driving to alert the driver in case of drowsiness. In further works, we want to explore other EEG datasets to investigate the feasibility of a large framework that can be applied to various fields, more the application of the model in a real-life application will be considered. Over the next years, we think that the use of EEG and ML models will be helpful to help in diagnosis and treatment, and to prevent accidents.

REFERENCES

[1] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[2] P. Tirupattur, Y. S. Rawat, C. Spampinato, and M. Shah, "ThoughtViz: Visualizing Human Thoughts Using Generative Adversarial Network," in *Proceedings of the 26th ACM international conference on Multimedia*, ser. MM '18. Association for Computing Machinery, 2018, pp. 950–958.

[3] V. Delvigne, H. Wannous, T. Dutoit, L. Ris, and J.-P. Vandeborre, "PhyDAA: Physiological Dataset Assessing Attention," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021, iEEE Transactions on Circuits and Systems for Video Technology.

[4] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, "Multi-channel EEG recordings during a sustained-attention driving task," *Nature Scientific Data*, vol. 6, no. 1, p. 19, 2019.

[5] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *Journal of Neural Engineering*, vol. 14, no. 2, 2017, iOP Publishing.

[6] Z. Gao, X. Wang, Y. Yang, C. Mu, Q. Cai, W. Dang, and S. Zuo, "EEG-Based Spatio–Temporal Convolutional Neural Network for Driver Fatigue Evaluation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2755–2763, 2019.

[7] B. Zou, M. Shen, X. Li, Y. Zheng, and L. Zhang, "EEG-based Driving Fatigue Detection during Operating the Steering Wheel Data Section*," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 248–251.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186.

[9] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, S.-W. Li, and H.-y. Lee, "Audio ALBERT: A Lite BERT for Self-supervised Learning of Audio Representation," *arXiv:2005.08575 [cs, eess]*, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[11] D. Merkx and S. L. Frank, "Comparing Transformers and RNNs on predicting human sentence processing data," *arXiv:2005.09471 [cs]*, 2020.

[12] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A Novel Bi-hemispheric Discrepancy Model for EEG Emotion Recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[13] P. Zhong, D. Wang, and C. Miao, "EEG-Based Emotion Recognition Using Regularized Graph Neural Networks," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[14] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, 2018, iOP Publishing.

[15] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks," in *4th International Conference on Learning Representations, ICLR*, 2016.

[16] K. H. Cheah, H. Nisar, V. V. Yap, and C.-Y. Lee, "Convolutional neural networks for classification of music-listening eeg: comparing 1d convolutional kernels with 2d kernels and cerebral laterality of musical influence," *Neural Computing and Applications*, vol. 32, no. 13, pp. 8867–8891, 2020.

[17] A. Ahangi, M. Karamnejad, N. Mohammadi, R. Ebrahimpour, and N. Bagheri, "Multiple classifier system for eeg signal classification with application to brain–computer interfaces," *Neural Computing and Applications*, vol. 23, no. 5, pp. 1319–1327, 2013.

[18] Y. Liu, Z. Lan, J. Cui, O. Sourina, and W. Müller-Wittig, "EEG-Based Cross-Subject Mental Fatigue Recognition," in *2019 International Conference on Cyberworlds (CW)*, 2019, pp. 247–252, iSSN: 2642-3596.

[19] Y. Liu, Z. Lan, J. Cui, O. Sourina, and W. Muller-Wittig, "Inter-subject transfer learning for EEG-based mental fatigue recognition," *Advanced Engineering Informatics*, vol. 46, 2020.

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

[21] T. Wen and Z. Zhang, "Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals," *IEEE Access*, vol. 6, pp. 25 399–25 410, 2018, iEEE Access.

[22] S. A. David, T. J. A. Machado, C. M. C. Inácio, and C. A. Valentim, "A combined measure to differentiate EEG signals using fractal dimension and MFDFA-Hurst," *Communications in Nonlinear Science and Numerical Simulation*, vol. 84, p. 105170, May 2020.

[23] B. Hjorth, "Eeg analysis based on time domain properties," *Electroencephalography and clinical neurophysiology*, vol. 29, no. 3, pp. 306–310, 1970.

[24] A. Erdamar, F. Duman, and S. Yetkin, "A wavelet and teager energy operator based method for automatic detection of k-complex in sleep eeg," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1284–1290, 2012.

[25] V. Delvigne, H. Wannous, J.-P. Vandeborre, L. Ris, and T. Dutoit, "Attention Estimation in Virtual Reality with EEG based Image Regression," in *IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2020, pp. 10–16.

[26] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "SST-EmotionNet: Spatial-Spectral-Temporal Based Attention 3D Dense Network for EEG Emotion Recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, 2020, pp. 2909–2917.

[27] Y. Yuan, K. Jia, F. Ma, G. Xun, Y. Wang, L. Su, and A. Zhang, "A hybrid self-attention deep learning framework for multivariate sleep stage classification," *BMC Bioinformatics*, vol. 20, no. 16, p. 586, 2019.

[28] C. Pernet, M. I. Garrido, A. Gramfort, N. Maurits, C. M. Michel, E. Pang, R. Salmelin, J. M. Schoffelen, P. A. Valdes-Sosa, and A. Puce, "Issues and recommendations from the OHBM COBIDAS MEEG committee for reproducible EEG and MEG research," *Nature Neuroscience*, vol. 23, no. 12, pp. 1473–1483, 2020.

[29] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[30] M. T. Martin, F. Pennini, and A. Plastino, "Fisher's information and the analysis of complex signals," *Physics Letters A*, vol. 256, no. 2, pp. 173–180, May 1999.

[31] Z. Mardi, S. N. M. Ashtiani, and M. Mikaili, "Eeg-based drowsiness detection for safe driving using chaotic features and statistical tests," *Journal of medical signals and sensors*, vol. 1, no. 2, p. 130, 2011.

[32] R. Hamila, J. Astola, F. A. Cheikh, M. Gabbouj, and M. Renfors, "Teager energy and the ambiguity function," *IEEE Transactions on Signal Processing*, vol. 47, no. 1, pp. 260–262, 1999.

[33] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, 2018, iOP Publishing.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, iSSN: 1063-6919.

[35] O. M. Bazanova and L. I. Aftanas, "Individual EEG Alpha Activity Analysis for Enhancement Neurofeedback Efficiency: Two Case Studies," *Journal of Neurotherapy*, vol. 14, no. 3, pp. 244–253, 2010.

[36] M. Arns, C. K. Conners, and H. C. Kraemer, "A decade of EEG Theta/Beta Ratio Research in ADHD: a meta-analysis," *Journal of Attention Disorders*, vol. 17, no. 5, pp. 374–383, Jul. 2013.

[37] B. J. Edelman, B. Baxter, and B. He, "Eeg source imaging enhances the decoding of complex right-hand motor imagery tasks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 1, pp. 4–14, 2015.

[38] V. V. Ogryzko, R. L. Schiltz, V. Russanova, B. H. Howard, and Y. Nakatani, "The Transcriptional Coactivators p300 and CBP Are Histone Acetyltransferases," *Cell*, vol. 87, no. 5, pp. 953–959, 1996.

[39] S. Bioulac, D. Purper-Ouakil, T. Ros, H. Blasco-Fontecilla, M. Prats, L. Mayaud, and D. Brandeis, "Personalized at-home neurofeedback compared with long-acting methylphenidate in an european non-

inferiority randomized trial in children with ADHD," *BMC Psychiatry*, vol. 19, no. 1, p. 237, 2019.

[40] F. Blume, J. Hudak, T. Dresler, A.-C. Ehlis, J. Kühnhausen, T. J. Renner, and C. Gawrilow, "NIRS-based neurofeedback training in a virtual reality classroom for children with attention-deficit/hyperactivity disorder: study protocol for a randomized controlled trial," *Trials*, vol. 18, Jan. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5259870/

[41] T. F. Collura and D. Siever, "Chapter 8 - Audio-visual entrainment in relation to mental health and EEG," in *Introduction to Quantitative EEG and Neurofeedback (Second Edition)*, T. H. Budzynski, H. K. Budzynski, J. R. Evans, and A. Abarbanel, Eds. San Diego: Academic Press, 2009, pp. 195–224.

[42] M. Behrmann, J. J. Geng, and S. Shomstein, "Parietal cortex and attention," *Current opinion in neurobiology*, vol. 14, no. 2, pp. 212–217, 2004.