2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Introducing Attention Mechanism for EEG Signals: Emotion Recognition with Vision Transformers

Arjun*, Aniket Singh Rajpoot* and Mahesh Raveendranatha Panicker, *Senior Member, IEEE*

***Abstract*— The accurate emotional assessment of humans can prove beneficial in health care, security investigations and human interaction. In contrast to emotion recognition from facial expressions which can prove to be inaccurate, analysis of electroencephalogram (EEG) activity is a more accurate representation of one's state of mind. With advancements in deep learning, various methods are being employed for this task. In this research, importance of attention mechanism in EEG signals is introduced through two vision transformer based methods for the classification of EEG signals on the basis of emotions. The first method utilizes 2-D images generated through continuous wavelet transform (CWT) of the raw EEG signals and the second method directly operates on the raw signal. The publicly available and widely accepted DEAP dataset has been utilized in this research for validating the proposed approaches. The proposed approaches report very high accuracies of 97% and 95.75% using CWT and 99.4% and 99.1% using raw signal for valence and arousal classifications respectively, which clearly highlights the significance of attention mechanism for EEG signals. The proposed methodology also ensures faster training and testing time which suits the clinical purposes.**

***Clinical Relevance*— This work establishes a highly accurate algorithm for emotion recognition using EEG signals which has potential applications in music-based therapy.**

I. Introduction

Emotions, the very essence of human beings, can be associated with thoughts, decision-making abilities and cognitive processes. Therefore, studies on emotional states can enhance current brain computer interface (BCI) systems which can be further employed in various applications such as implementing therapies for disorders like autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and anxiety disorder [1]. Due to such important applications, recognition and analysis of emotional states have become an important research area in the fields of medical science, neuroscience, cognitive science and brain driven artificial intelligence. Several methods have been developed for emotion recognition which includes the use of both physiological and non-physiological signals. Non-physiological signals include facial expressions, voice signals, body gestures while physiological signals include EEG, ECG signals and many more. Using non-physiological signals is relatively easy and does not require any special equipment, but an individual can forge such signals and are therefore not considered as a true reflection of one's emotional state. In contrast, physiological signals are beyond one's control and therefore more suitable for the given task [7].

Various studies have been done in the past that have specifically handled emotion recognition through physiological signals as in [2-9]. Algorithms using power spectral density (PSD) features with Naive Bayes classifiers [2, 3], PSD and Statistical features with Ontological models [4], Deep belief network (DBN) based features with support vector machine classifier [5], power spectral and statistical features with neural networks (NN) [6], features extracted using LP-1D-CNN model with SoftMax as a classifier [7], Pearson Correlation Coefficient features with Deep Neural Network and Sparse Autoencoder architecture as a classifier [8], and raw EEG 1D time signals directly used with MMResLSTM as a classifier [9] are a few of them. In most of the approaches [2-9], emotional states, which are ideally discretized into numerous states such as joy, fear, anger, happiness, surprise etc., are broadly classified into two basic meaningful dimensions: valence and arousal[18]. The valence dimension determines the positive or negative effects of the emotion and the arousal dimension determines the intensity of it as shown in Fig 1.

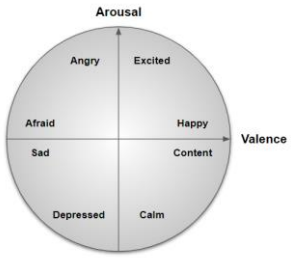


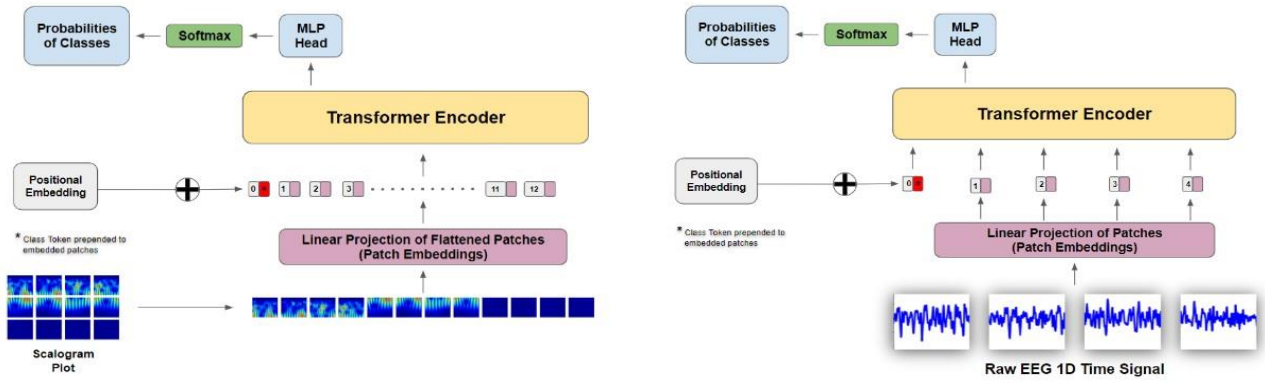
Fig. 1 Illustration of Valence and Arousal Theory

However, it has to be remembered that, tasks like emotion recognition occur over a period of seconds and are not an instantaneous response which happens over a period of milliseconds. As a few seconds of time is a significant amount of data for EEG, there might be connections between an impulse occurred between a brief period of time. In such cases it will be good if the model employed for emotion classification considers events that happened far in the past also. Architectures such as Convolutional Neural Networks (CNN) and Long-short-term-memory (LSTM) may not be able to consider this long-term dependence. CNN's are localized networks as determined by the kernel size and respective strides, whereas LSTMs do not have good memory

*Equal contributions. This work was supported by Indian Institute of Technology (IIT) Palakkad, India.

Arjun (email: beniwal.arjun1@gmail.com) and Mahesh Raveendranatha Panicker (email: mahesh@iitpkd.ac.in) are with Electrical Engineering, IIT Palakkad, India. Aniket Singh Rajpoot (email: aniket161200@gmail.com) is with Computer Science and Engineering, IIT Palakkad, India.

The codes are available at https://github.com/AniketRajpoot/Emotion-Recognition-Transformers



(a) (b)

Fig. 2: The proposed ViT based architecture: (a) with input as images generated using CWT, (b) with input as Raw EEG

retaining capability due to the forgetting factor [10, 11]. On the other hand, the ability to model dependencies without having the constraint of far distances in the sequence is basically the very core of attention mechanism in transformer networks [12, 13]. Transformers [10] which are based on the self-attention mechanism, have been very widely accepted in natural language processing (NLP) because of this. At a high level, the model goes through every vector where self-attention enables it to look at other parts of the input sequence which can help in the better encoding of the vector. A transformer network is a stack of these attention layers with some residuals connections. Transformers have the capability to retain as much information as the memory limits, and establish a relationship between what has occurred in the past and what is happening now. LSTM's and CNN's model their positions in relative terms whereas transformers rely on absolute position representations of the input (The positional embedding and it is permutation invariant) [10, 11].

In this research, the variant of the transformer called Vision Transformer (ViT) [11] which was made specifically for images, has been adapted to emotion detection in EEGs. The reason for choosing ViT is to employ time-frequency images as generated by wavelet transforms, which takes into account, the localized variations in frequency. However direct application of ViT on the raw EEG signal gave a significant improvement in accuracy as evident from the results when compared to time-frequency images. This clearly shows two aspects 1) the significance of attention mechanism for EEG signals and 2) the need of a proper encoding scheme. To the best of our knowledge, this is the first attempt of employing ViT for EEG signal analysis and also the first effort towards identifying the significance of attention in EEG signals. One of the biggest advantages of the simple setup of ViT is that they are scalable and efficient.

II. Proposed Methodology

In this section, the proposed approach of ViT for CWT images and the raw EEG signal is explained in detail.

A. Model Architecture

The architecture for ViT [11] closely resembles that of the vanilla transformer [10]. NLP transformers have token embeddings, meaning that it receives 1D input with a known dictionary size as input. However, for 2D input as in the case of ViT, the image is divided into a sequence of flattened 2D fixed size image patches which act as tokens. Therefore, an image of size $x \in \mathbb{R}^{H\times W\times C}$ is divided into sequences of patches of size $x \in \mathbb{R}^{N\times(P^2\times C)}$ where $N = HW/P^2$ and P is the selected patch size. Finally, before passing the obtained patches to the transformer it is passed through a trainable linear projection layer as in (1) [11] for getting the final patch embeddings(z_0). ViT uses these patch embeddings so that there is no constraint of a certain vocab like in NLP transformers.

$$z_0 = [x_{class};x_p{}^1E;x_p{}^2E;\dots;x_p{}^nE]+E_{pos},$$

$$E\in R^{D\times(P^2.C)},E_{pos}\in R^{D\times(N+1)} \quad (1)$$

$$z_l{}'=MSA(LN(z_{l-1}))+z_{l-1},\ l=1,2,\dots,L \quad (2)$$

$$z_l{}'=MLP(LN(z_l{}'))+z_l{}',\ l=1,2,\dots,L \quad (3)$$

Similar to Bidirectional Encoder Representations from Transformers (BERT) [14] architecture, a learnable class token embedding is prepended to the patch embeddings. Positional embeddings(E_{pos}) are also added to these patch embeddings for introducing positional information of the tokens in the sequence. The transformer model contains alternating layers of Multiheaded Self-Attention (MSA) and MLP (2 layers with Gaussian Error Linear Unit (GELU) non-linearity) blocks (as in (2), (3)) with a layer normalization(LN) before every block and residual connections after every block [15, 16].

B. Feature Extraction

In the proposed ViT based EEG classifier network, the input data to the ViT is considered in 2 ways i.e., raw EEG signal and the image generated through CWT. The architecture of the proposed approach is as shown in Figures 2a and 2b. The application of Wavelet transforms in EEG has been very popular due to their compression and time-frequency localization capabilities [17]. The choice of mother wavelet used is an important aspect based on its compatibility with the time signal. As studied in [17] EEG signals are most compatible with near symmetric and orthogonal mother wavelets like sym24, db4, coif5. In this research work db4 and coif5 mother wavelets have been employed for generating the

images to be employed as the input to the ViT. As part of the ablation, experiments other compressed representations such as auto encoder [8] have been tried instead of the CWT based images, however the results were not encouraging.

III. Results and Analysis

In this section, the details of the dataset employed for the analysis of the proposed approaches and also the analysis results are presented.

A. Dataset Description

The proposed method was validated with the widely used DEAP [2] dataset. In this dataset EEG and peripheral physiological signals of 32 participants were recorded. Each participant in this dataset watched 40 one-minute music videos and simultaneously their EEG recordings were taken at a sampling rate of 512 Hz with 32 channels which was later down sampled to 128 Hz and band pass filtered to 4 – 45 Hz. Each video was rated by the participants on the basis of mainly valence, arousal, liking and dominance in a range of 1-9. With the DEAP dataset, many classes can be extracted from labels by dividing them equally. In the proposed work, two class labels for valence and arousal are adopted.

B. Training

The 60 second recording of each video was broken down into non-overlapping smaller n-sized samples *(n = 6, 15, 20, 30 in seconds)* respectively as responses like emotion develop over a few seconds of time period and thus breaking down the video into these sizes would help us to focus on each emotion development properly. The dataset with the above sized samples with all the 32 channels was fed to the ViT model and trained through the following two ways, which can be seen in the Fig 2.

- **Image generated through CWT:** The n-sized 32 channel sample was transformed using CWT with 48 scales and employing db4 and coif5 mother wavelets. The scalogram images generated as part of the 48 scale CWT are then fed to the ViT, in which patch embedding, with a shape of *[patchsize, patchsize]* is applied to it. The flattened patches are mapped to D dimensions with a trainable linear projection layer (as in (1)). Now, a class token is prepended to the output received from the trainable linear projection layer. Finally, positional embeddings are added to the patch embeddings and it is transferred to the transformer encoder.

TABLE I: MEAN Classification accuracy of Valence through CWT

Wavelet	6 sec	15 sec	20 sec	30 sec
db4	95.7%	92.7%	91.15%	87.5%
coif5	97%	93.75%	92%	88%

TABLE II: MEAN Classification accuracy of Arousal through CWT

Wavelet	6 sec	15 sec	20 sec	30 sec
db4	95.5%	93.9%	92.5%	85.15%
coif5	95.75%	94.4%	92.9%	89.45%

TABLE III: MEAN Classification accuracy of valence and arousal through raw EEG signal

Labels	6 sec	15 sec	20 sec	30 sec
Valence	99.4%	99.2%	97.5%	92%
Arousal	99.1%	99.2%	98%	90.5%

- **Raw EEG Signal:** In this case, instead of any transformation or encoding, the raw 32 channel EEG signal (preprocessed with the 4 – 45 Hz bandpass filters as part of the DEAP dataset) is directly sent to the ViT, as shown in Fig. 2b. Since the raw EEG signal is a 1D time signal, the patch embedding is applied with a shape of *[1, patchsize]*. Also, as the patches are already flattened in this case, they are directly mapped to D dimensions with a trainable linear projection. Similarly, a class token is prepended to it followed by the addition of positional embeddings and finally transferring to the transformer encoder.

The output from the transformer encoder, in both the CWT images and raw EEG signal-based models, is passed through a MLP head layer where it is mapped to the number of classes. Then a SoftMax layer followed by an ArgMax layer is applied for the getting the class with maximum probability. A 6-layered transformer with an embedding dimension of 512 and 8 heads for MSA was used for training. As compared to its counterparts in NLP, the size and memory usage of this transformer is 2-3x times smaller which results in faster training and testing time [11]. In this work, the implementation was done on Python 3.7.10 and TensorFlow 2.5.0. The learning rate is set as 0.00001.

C. Results

As discussed in section III.A, to verify the effectiveness of the proposed approaches, experiments were carried out on the publicly available DEAP dataset [2]. The dataset was divided,

TABLE IV: MEAN Classification accuracy of valence and arousal through raw EEG signal

Research	Features	Classifier	Valence	Arousal
Koelstra et al. [2], 2012	PSD	Gaussian Naive Bayes	57.6%	62.0%
Chung and Yoon [3], 2012	PSD	Naive Bayes	66.6%	66.4%
Zhang et al. [4], 2013	PSD, Statistical features	Ontological model	75.19%	81.74%
Liu et al. [5], 2016	DBN based features	SVM	85.2%	80.5%
Yin et al. [6], 2017	PSD, Statistical features	Neural networks	83.04%	84.18%
Emad-ul-Haq Qazi et al [7], 2019	Features extracted using LP-1D-CNN model	SoftMax	**98.43%**	**97.65%**
Junxiu Liu et al. [8], 2020	PCC	Deep neural network and sparse autoencoder	89.49%	92.86%
Jiaxin Ma et al. [9], 2019	Raw EEG 1-D time signal	MMResLSTM	92.87±2.11	92.30±1.55
Proposed Approaches	Image generated through CWT	Vision Transformer	**97%**	**95.75%**
	Raw EEG 1-D time signal	Vision Transformer	**99.4%**	**99.1%**

such that 80% of the data went to the training set and the remaining 20% to testing set.

- **Image generated through CWT:** The results of the Images generated through CWT can be seen in Table I and II. As shown, the scalograms formed by the 6 second sized samples performed significantly better than the 15, 20 and 30 second sized samples. This clearly shows the significant localized behavior of the EEG signals and the importance of a model which can take localized regions of EEG for further processing.
- **Raw EEG Signal:** The results of the raw EEG signal experiment can be seen in Table III. In this case, as can be seen, 6 and 15 second sized samples performed significantly better than the 20 and 30 sized samples. More importantly, on comparing Tables I, II and III, it is evident that the raw EEG signal-based approach surprisingly performs much better than the CWT based approach. This could be attributed to the fact that, the EEG signals being random and the emotion content being local, a transformation of EEG signal is not needed (or need to be applied carefully) in the case when attention approaches like transformers [11] are employed. A detailed analysis of the same will be done as part of the future study.

The proposed approaches are also compared in a comprehensive fashion with most of the well-established approaches in literature and the results are reported in Table IV. From Table IV, it can be seen that the proposed ViT based method outperformed all the recent-related state-of-the-art studies documented in literature. The main reason for getting good results through ViT could be attributed to the attention-based mechanism. Through the multi-headed attention-based mechanism, the model is able to capture and remember the development of emotion through time in a much better and faster way than what CNN's and LSTM's or hand-crafted machine learning algorithms can do. It can also be noted that, the results presented in this work agree to the observation reported by most of the established research works related to classification of emotion through EEG signals, that the smaller sized samples perform better than the longer sized samples.

IV. Conclusions

In this paper, we investigated two experimental setups i.e., image generated through CWT and Raw Signal for EEG based emotion recognition with Vision Transformers (ViT). The ViT yielded good results with the publicly available DEAP dataset with an accuracy of 97% and 95.75% for valence and arousal in the Image Formed through CWT experiment with Coif5 mother wavelet. On the other hand, an accuracy of 99.4% and 99.1% for valence and arousal in the Raw EEG signal experiment, thereby outperforming the existing state-of-the-art methods. One of the main reasons for the exceptional performance of ViT is the attention-based mechanism, due to which it is able to capture and retain more relevant information than the conventional CNNs' and LSTMs'. Both the experiments conducted also confirmed that smaller sized samples are more optimal for capturing the emotions, as they yield a higher classification accuracy than others. Furthermore, ViTs are more computationally faster than other neural networks for similar tasks, which makes them more suitable for real time analysis tasks. Future work involves a thorough comparison of various compression/ encoding schemes as input to ViT as well as an approach to identify the most influential EEG channels and also quantify the influence of the time-segment which resulted in the highest attention score particularly in raw EEG signal experiments.

References

[1] Gantayat, S. Sekhar, and S. Lenka, "Study of Algorithms and Methods on Emotion Detection from Facial Expressions: A Review from Past Research," *Communication Software and Networks*, pp. 231- 44, 2021

[2] S. Koelstra et al., "DEAP: A Database for Emotion Analysis; Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18-31, June 2011.

[3] S. Y. Chung and H. J. Yoon, "Affective classification using Bayesian classifier and supervised learning," in *12th International conference on Control, Automation and Systems*, pp. 1768-1771, Jeju Island, 2012.

[4] X. Zhang, B. Hu, J. Chen and P. Moore, "Ontology-based context modeling for emotion recognition in an intelligent Web", *World Wide Web*, vol. 16, no. 4, pp. 497-513, 2013.

[5] Liu, Wei, W.L. Zheng, and B. L. Lu, "Emotion recognition using multimodal deep learning," in *International conference on neural information processing*. Springer, Cham, 2016.

[6] Z. Yin, M. Zhao, Y. Wang, J. Yang and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 93-110, Mar. 2017.

[7] Qazi, E. U. Haq, M. Hussain, H. AboAlsamh, and I. Ullah, "Automatic Emotion Recognition (AER) System based on Two-Level Ensemble of Lightweight Deep CNN Models," *arXiv preprint*, arXiv:1904.13234, 2019.

[8] J. Liu, G. Wu, Y. Luo, et. al., "EEG based Emotion Classification Using Deep Neural Network and Sparse Autoencoder," *Frontiers in Systems Neuroscience,* vol. 14, pp. 43, 2020.

[9] J. Ma, H. Tang, W.-L. Zheng and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network", in *Proc. 27th ACM Int. Conf. Multimedia*, pp. 176-183, Oct. 2019.

[10] A. Vaswani, N. Shazeer and N. Parmar, et.al, "Attention is all you need", in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, pp. 5999-6009, 2017.

[11] Dosovitskiy, Alexey, et. al., "An image is worth 16x16 words: Transformers for image recognition at scale", *arXiv preprint* arXiv:2010.11929, 2020.

[12] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate", in *International Conference on Learning Representations*, 2015.

[13] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," *arXiv preprint,* arXiv:1702.00887, 2017.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *in NAACL-HLT*, pp. 4171–4186, 2019.

[15] Q. Wang, B. Li, X. Tong, et. al., "Learning deep transformer models for machine translation", in *ACL*, pp. 1810-1822, 2019.

[16] A. Baevski and M. Auli, "Adaptive input representations for neural language modeling", in *ICLR*, May 2019.

[17] M. I. AI-Kadi, M. B. I. Reaz and M. A. Mohd Ali, "Compatibility of mother wavelet functions with the electroencephalographic signal", *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences IECBES 2012*, pp. 113-117, 2012.

[18] J.A. Russell, "A Circumplex Model of Affect", *J. Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.