2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10781700

Transformer-Based Emotion Recognition with EEG

Kulin Patel[1], Farshad Safavi[1], Rajarathnam Chandramouli[1,2], and Ramana Vinjamuri[1], *IEEE Senior Member*
[1]Vinjamuri Lab, University of Maryland Baltimore County, Baltimore, MD, USA. [2]Spectronn, NJ, USA.
Email: rvinjam1@umbc.edu

***Abstract*— Emotion recognition via electroencephalography (EEG) has emerged as a pivotal domain in biomedical signal processing, offering valuable insights into affective states. This paper presents a novel approach utilizing a tailored Transformer-based model to predict valence and arousal levels from EEG signals. Diverging from traditional Transformers handling singular sequential data, our model adeptly accommodates multiple EEG channels concurrently, enhancing its ability to discern intricate temporal patterns across the brain. The modified Transformer architecture enables comprehensive exploration of spatiotemporal dynamics linked with emotional states. Demonstrating robust performance, the model achieves mean accuracies of 92.66% for valence and 91.17% for arousal prediction, validated through 10-fold cross-validation across subjects on the DEAP dataset. Trained for subject-specific analysis, our methodology offers promising avenues for enhancing understanding and applications in emotion recognition through EEG. This research contributes to a broader discourse in biomedical signal processing, paving the way for refined methodologies in decoding neural correlates of emotions with implications across various domains including brain-computer interfaces, and human-robot interaction.**

Keywords*— *EEG, Transformers, emotion recognition, brain-computer interfaces, human-robot interaction

I. Introduction

Emotion recognition through electroencephalography (EEG) has emerged as a pivotal domain in affective computing, holding profound implications for brain-computer interfaces, human-robot interaction and other healthcare applications. The intrinsic link between neural activity and emotional states makes EEG a valuable modality for decoding underlying emotional cues. This paper delves into the EEG-based emotion recognition, addressing the challenges posed by noisy, non-linear, and non-stationary nature of EEG signals.

EEG, being a direct measure of brain activity, provides a unique window into the dynamics of emotional states. Its non-invasiveness and temporal precision make it an ideal candidate for real-time emotion recognition, with applications spanning human-robot interaction, healthcare diagnostics, and personalized affective computing systems. The ability to decode emotional states from EEG signals unlocks a myriad of possibilities for enhancing human-machine collaboration and understanding.

While EEG-based emotion recognition has gained traction, existing methodologies face inherent challenges. Notably, the intricate nature of EEG signals demands intelligent frameworks capable of providing high accuracy in emotion recognition. Various approaches, including classical machine learning models, have been explored to tackle these challenges. However, interpretability, noise resilience, manual feature extraction, and subject-specific adaptability remain ongoing concerns.

The emergence of deep learning models has had a significant impact on EEG classification, moving beyond traditional machine learning methods. Deep learning techniques utilize comprehensive artificial neural networks instead of conventional methods that require manual feature extraction. Convolutional neural networks (CNN) [1], [2] Long short-term memory (LSTM) [3], [4], and hybrid CNN models are widely used deep learning techniques for analyzing EEG signals.

While deep learning methods like CNN and LSTM have been applied for EEG-based emotion recognition, inherent challenges persist. CNNs struggle with capturing long-range dependencies in sequential data, that are crucial for understanding the temporal dynamics of emotional states in EEG signals. On the other hand, LSTMs, while adept at handling sequential data, may face difficulties in managing the complex and non-linear nature of EEG signals. These challenges underscore the need for a more adaptive and efficient approach.

The emergence of Transformer models [5], known for their ability in handling sequential data with a self-attention mechanism, present a compelling solution. By addressing the limitations of traditional models, Transformers offer a unique opportunity to enhance the accuracy and interpretability of emotion recognition from EEG signals. Given their outstanding performance in natural language processing (NLP) and computer vision (CV), employing Transformers to process entire EEG signals [6] presents a viable and effective approach. Consequently, Transformers designed for EEG applications have been developed for various uses, including emotion recognition. However, it's noteworthy that not all existing models harness the full potential of multi-channel EEG data effectively. For instance, ERTNet [7] and STS-Transformer[8], despite incorporating a Transformer-based model, fall short in delivering optimal results. One prominent limitation lies in the absence of an equivalent mechanism to word embedding, as seen in natural language processing (NLP). In EEG-based emotion recognition, this can hinder the models' ability to capture local patterns effectively. This emphasizes the crucial need for a model specifically tailored to leverage all EEG channels comprehensively, ensuring the comprehensive capture of intricate temporal patterns distributed across the brain.

This paper introduces an innovative approach to EEG-based emotion recognition, leveraging a modified Transformer-based model. Departing from conventional Transformer model which can process one sequence at a time, our model is tailored to handle multiple EEG channels simultaneously, enriching its

Research supported by NSF CAREER Award (HCC-2053498).

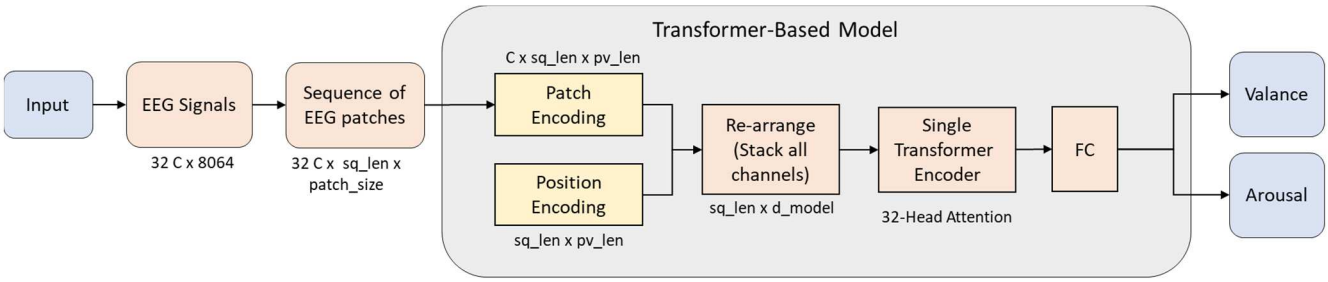


Fig1: Proposed Model Architecture (Here C = Channels = 32, sq_len = Sequence length , patch_size = length of each patch, pv_len = Patch vector length of encoded patch, d_model = C x pv_len, sq_len x patch_size = 8064)

capacity to discern intricate temporal patterns across the brain. The modified Transformer architecture facilitates a comprehensive understanding of spatiotemporal dynamics associated with emotional states, promising enhanced accuracy and interpretability.

The subsequent sections detailed the methodology, model architecture, training procedures, and evaluation metrics, providing a holistic understanding of the proposed approach and its implications for advancing EEG-based emotion recognition.

II. Methods

A. EEG Data Source

1) DEAP Dataset Overview

The study utilizes the DEAP dataset [9], a comprehensive multimodal resource for the exploration of human affective states. The dataset captured electroencephalogram (EEG) and peripheral physiological signals from 32 participants, each exposed to 40 one-minute excerpts of music videos. These participants provided subjective ratings on arousal, valence, like/dislike, dominance, and familiarity, ranging from 0 to 9 for each dimension.

TABLE I. DEAP DATASET

	Array Shape	Array Content
Data	40 × 32 × 8064	Video × Channels × Data
Labels	40 × 4	Videos × Label

2) Data Label Transformation

For the purpose of binary classification in predicting valence and arousal, a label transformation was applied. Ratings for valence and arousal were scaled to binary values, where a rating of 5 or higher was mapped to 1, indicating a positive response, and ratings below 5 were mapped to 0, indicating a negative response.

B. Model Architecture

Our proposed model architecture is based on original Transformer [5], a deep learning architecture originally designed for sequence-to-sequence tasks. Departing from the conventional use of Transformers for single sequence/sentence, our model is adapted to handle multi-channel EEG data effectively. The architecture is optimized to capture spatiotemporal patterns across 32 EEG channels, offering a more nuanced understanding of neural dynamics associated with emotional states.

The architecture comprises three key blocks: Patch Encoding, Position Encoding, and a Single Transformer Encoder with 32 heads, corresponding to the number of EEG channels. The output from the Transformer encoder is fed into a fully connected layer for final predictions.

1) Patch Encoding

In our model architecture, Patch Encoding plays a pivotal role akin to word embedding in NLP. This critical step involves encoding each patch of EEG data to capture underlying patterns effectively.

Initially organized as a 32 × 8064 matrix, the EEG signals undergo a patch-based transformation. Through systematic division into patches of size *patch_size* , the matrix is reshaped into dimensions of 32 × *sq_len* × *patch_size*. This strategic patching ensures the precise capture of local patterns inherent in EEG data. Subsequently, these patches are encoded via a linear layer, producing patch vectors whose lengths *pv_len*, correspond to the output size of the linear layer. This transformation ensures that intricate temporal dynamics across different channels are accurately represented. Visual representation of this process is depicted in Figure 1, elucidating the sequential flow of patch encoding within our model architecture.

2) Position Encoding

Post patch encoding, the shape transforms to 32 × *sq_len* × *pv_len*. To integrate temporal information into the model, a crucial step involves position encoding - a practice inspired by the conventional techniques observed in Transformers [5], infusing the data with crucial temporal context.

The result of this process mirrors the output of the patch encoding phase. After adding Position Encodings and Patch Encodings, the result undergoes a reorganization into a 2D representation, precisely a matrix with dimensions of *sq_len* × (32 × *pv_len*). This matrix is fed into the subsequent Transformer encoder, enriching the model's understanding of the temporal dynamics inherent in the EEG data.

3) Single Transformer Encoder with 32 Heads

The heart of the model lies in the Transformer encoder, a critical component adept at processing multi-channel EEG data. The encoder is designed to treat each channel's information independently, fostering a nuanced understanding of the diverse neural activities across the 32 channels. While retaining the fundamentals of the original Transformer model, we have streamlined the encoder architecture to utilize only one block, in contrast to the six blocks in the original design.

Notably, the Transformer encoder incorporates 32 heads, enabling simultaneous attention to various aspects of the input data. This intricate attention mechanism contributes significantly to the model's capacity to discern complex spatiotemporal patterns in EEG data.

4) Output and Fully Connected Layer (FC):

The output from the Transformer encoder is seamlessly directed to a fully connected layer (FC). This network is designed to predict two values: one for valence and another for arousal.

Fig. 1 represents the model architecture, illustrating the sequential flow from patch encoding to the final output through the Transformer encoder and FC. The subsequent sections will delve into specifics of training, hyperparameters, and evaluation metrics, offering a holistic understanding of the model's performance in predicting valence and arousal.

C. Training Procedure

1) Subject-Specific Analysis

To enhance the model's adaptability and account for inter-subject variability, a subject-specific analysis was employed during the training process. This approach involves training the model individually for each subject in the dataset, ensuring that the model can capture subject-specific patterns in EEG data.

2) 10-fold Cross-Validation

The model's robustness and generalizability were assessed using 10-fold cross-validation. The dataset was divided into 10 subsets, with the model trained and validated iteratively on different combinations of these subsets. This cross-validation strategy helps mitigate overfitting and provides a more reliable estimate of the model's performance.

3) Implementation and Hyperparameters used

We trained the proposed method on an NVIDIA GeForce 4070 Ti GPU. The hyperparameters of the proposed model were tuned to optimize performance, as outlined in Table II.

TABLE II HYPERPARAMETERS USED FOR THE TRAINING THE MODEL

Hyperparameter	Type/Value
Batch Size	4
Learning Rate	0.0001
Encoder Dropout	0.1
Epochs for each fold	1000
Patch Size	8
Attention heads in encoder	32
Optimizer	Adam

These hyperparameters were chosen through iterative experimentation to strike a balance between model convergence and generalization. The small batch size, low learning rate, and dropout regularization contribute to stable training, while the choice of the Adam optimizer and mean squared error loss function aligns with the nature of the regression task for valence and arousal prediction.

D. Evaluation Metrics

1) Subject-specific Accuracy:

Subject-specific accuracy was computed as the mean accuracy across the 10-fold cross-validation for each individual subject. The formula for subject-specific accuracy is given by:

$$Accuracy_{subject} = \left(\frac{1}{10}\right) \Sigma \frac{(TP_i + TN_i)}{(TP_i + TN_i + FP_i + FN_i)}$$

Where:

TP_i , TN_i = True Positives/Negatives for subject i

FP_i, FN_i = False Positives/ Negatives for subject i

2) Mean Accuracy Across Subjects:

The overall mean accuracy across all subjects was then calculated as the average of the subject-specific accuracies. Additionally, the standard deviation of these mean accuracies provided insights into the consistency of model performance across different subjects.

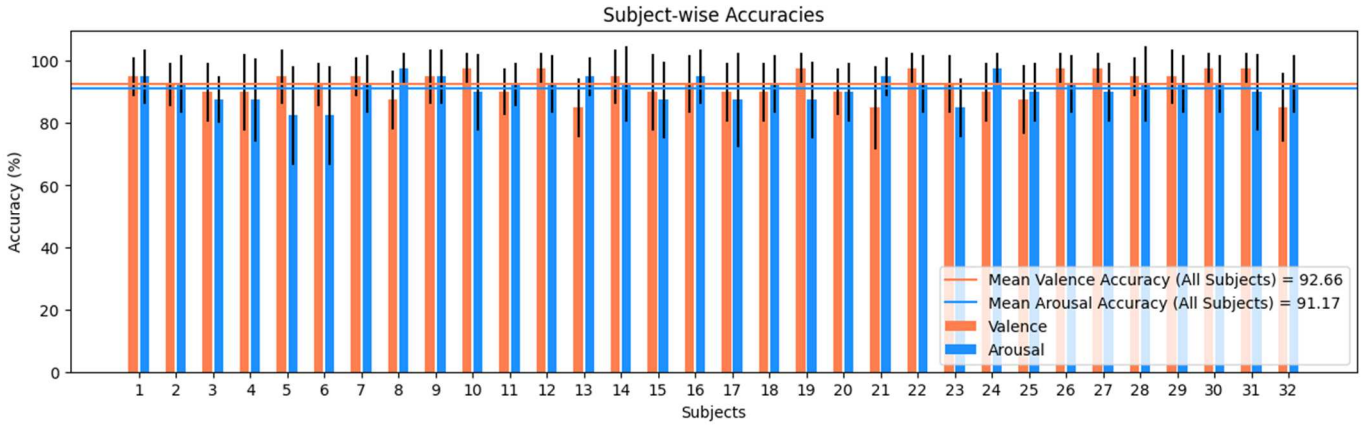
III. RESULTS

The comparative analysis aimed to assess the effectiveness of the proposed model against existing methodologies in EEG-based emotion recognition, with mean accuracies presented in Table III for valence and arousal predictions across various models tested. Fig. 2 illustrates these values for all subjects for the proposed method.

TABLE III COMPARISON OF MEAN ACCURACIES FOR VARIOUS METHORDS FOR VALANCE AND AROUSAL.

Acc (%)	Year	Valence %	Arousal %
LSTM	2017	84.75	82.16
CNN+HMM	2020	79.77 ± 0.61	83.09 ± 0.84
HOLO-FM	2021	76.61 ± 2.13	77.72 ± 2.87
DBCN	2022	87 ± 4.5	90.93 ± 3.9
ERTNet	2024	73.31	80.99
STS-Transformer	2024	89.86	86.83
Proposed Method	**2024**	**92.66 ± 3.99**	**91.17 ± 3.67**

The LSTM model [3] achieved good accuracies of 85.45% for valence and 85.65% for arousal, showcasing its proficiency in handling sequential data. However, it is worth noting that LSTM models might struggle with capturing long-range dependencies in EEG signals, potentially limiting their ability to discern intricate temporal patterns. The CNN+HMM hybrid model [10], incorporating convolutional neural networks and hidden Markov models, exhibited accuracies of 79.77% (valence) and 83.09% (arousal). Despite its capabilities, the reliance on manual feature extraction and the inherent challenges of hidden Markov models might impede its adaptability and noise resilience. HOLO-FM [11], utilizing holographic feature maps, demonstrated accuracies of 76.61% (valence) and 77.72% (arousal). While innovative, this approach may face limitations in capturing the complexity of emotional states due to its reliance on feature maps. DBCN [2], employing dilated bottleneck-based convolutional neural networks, exhibited high accuracies of 87% (valence) and 90.93% (arousal). However, the interpretability of such complex architectures might be



compromised, and subject-specific adaptability could be challenging.

ERTNet [7] and STS-Transformer [8], both Transformer-based approaches lacking a word embedding equivalent, as we have patch encoding, showed accuracies of 73.31% and 80.99% for ERTNet, and 89.86% and 86.83% for STS-Transformer, in valence and arousal prediction, respectively. The absence of mechanisms equivalent to word embedding and patch encoding might hinder these models in effectively capturing local patterns and spatiotemporal dynamics in EEG data. In contrast, the proposed model, incorporating patch encoding and handling multi-channel EEG, with each channel having a separate head, achieved mean accuracies of 92.66% (valence) and 91.17% (arousal). This design choice enables the model to comprehensively discern intricate temporal patterns across the brain, demonstrating enhanced accuracy and interpretability.

IV. Discussion and Conclusion

The presented results underscore the effectiveness of the proposed method in EEG-based emotion recognition by valence and arousal prediction. Leveraging a modified Transformer-based architecture tailored for multi-channel EEG data, our approach enabled a comprehensive understanding of spatiotemporal dynamics associated with emotional states. The competitive accuracy achieved, despite slight variability, highlights the robustness and adaptability of our model across subjects. This positions our method as a promising tool for advancing EEG-based emotion recognition, with implications for affective computing and human-machine interaction [12]. However, it's essential to acknowledge the limitations observed in our study. While our model demonstrated better performance, there is still room for improvement in accuracy, particularly in addressing the variability observed across subjects. In the near future we will refine our model to enhance accuracy and reliability, via multilayer preprocessing. Furthermore, applying our method to other datasets and real-world scenarios could validate its generalizability and contribute to the development of robust and versatile emotion recognition systems, thus continue to improve EEG-based emotion recognition.

Acknowledgment

The authors would like to thank the University of Maryland Baltimore County (UMBC) START award, UMBC Computer Science and Electrical Engineering Department, UMBC bwtech Maryland New Venture Fellowship Program, NSF CAREER award and NSF I-CORPs award for their support.

References

[1] Z. Tang, C. Li, and S. Sun, "Single-trial EEG classification of motor imagery using deep convolutional neural networks," *Optik (Stuttg)*, vol. 130, pp. 11–18, Feb. 2017, doi: 10.1016/J.IJLEO.2016.10.117.

[2] Q. Yao, H. Gu, S. Wang, and X. Li, "A Feature-Fused Convolutional Neural Network for Emotion Recognition from Multichannel EEG Signals," *IEEE Sens J*, vol. 22, no. 12, pp. 11954–11964, Jun. 2022, doi: 10.1109/JSEN.2022.3172133.

[3] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion Recognition based on EEG using LSTM Recurrent Neural Network," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017, Accessed: Feb. 01, 2024. [Online]. Available: www.ijacsa.thesai.org

[4] X. Zheng and W. Chen, "An Attention-based Bi-LSTM Method for Visual Object Classification via EEG," *Biomed Signal Process Control*, vol. 63, p. 102174, Jan. 2021, doi: 10.1016/J.BSPC.2020.102174.

[5] A. Vaswani *et al.*, "Attention is All you Need," *Adv Neural Inf Process Syst*, vol. 30, 2017.

[6] A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker, "Introducing Attention Mechanism for EEG Signals: Emotion Recognition with Vision Transformers," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 5723–5726, 2021, doi: 10.1109/EMBC46164.2021.9629837.

[7] R. Liu *et al.*, "ERTNet: an interpretable transformer-based framework for EEG emotion recognition," *Front Neurosci*, vol. 18, p. 1320645, Jan. 2024, doi: 10.3389/FNINS.2024.1320645.

[8] W. Zheng and B. Pan, "A spatiotemporal symmetrical transformer structure for EEG emotion recognition," *Biomed Signal Process Control*, vol. 87, p. 105487, Jan. 2024, doi: 10.1016/J.BSPC.2023.105487.

[9] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans Affect Comput*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[10] Q. Zhong, Y. Zhu, D. Cai, L. Xiao, and H. Zhang, "Electroencephalogram Access for Emotion Recognition Based on a Deep Hybrid Network," *Front Hum Neurosci*, vol. 14, p. 589001, Dec. 2020, doi: 10.3389/FNHUM.2020.589001/BIBTEX.

[11] A. Topic and M. Russo, "Emotion recognition based on EEG feature maps through deep learning network," *Engineering Science and Technology, an International Journal*, vol. 24, no. 6, pp. 1442–1454, Dec. 2021, doi: 10.1016/J.JESTCH.2021.03.012.

[12] F. Safavi, K. Patel, and R. K. Vinjamuri, "Towards Efficient Deep Learning Models for Facial Expression Recognition using Transformers," pp. 1–4, Dec. 2023, doi: 10.1109/BSN58485.2023.10331041.