TWEN: EEG Emotion Recognition Model Based on Weakly Supervised Learning Framework with Two-Phase Multitask Autoencoder

Taewan Kim
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Seoul, South Korea
23102324@seoultech.ac.kr

ChangGyun Jin
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Seoul, South Korea
jcg6074@seoultech.ac.kr

Seong-Eun Kim
Department of Applied Artificial Intelligence
Seoul National University of Science and Technology
Seoul, South Korea
sekim@seoultech.ac.kr

Abstract— **Emotions significantly influence human cognition, behavior, and social interactions, making accurate recognition essential in Human-Computer Interaction (HCI) applications. This study addresses challenges in EEG-based emotion recognition, particularly inter-subject variability and label noise, which hinder the development of robust and generalized models. We propose a robust Two-phase Weakly Supervised Emotion Network (TWEN), a novel deep learning model designed to enhance emotion recognition. TWEN incorporates a Two-phase Multitask Autoencoder to mitigate inter-subject variability and a Top-k Selection method to reduce label noise. The model captures both local and global temporal features of EEG signals through an innovative fusion of attention mechanisms, ensuring accurate classification of emotions over varying durations. Evaluations on the THU-EP dataset demonstrate that TWEN outperforms state-of-the-art models, achieving a classification accuracy of 60.8%, with a standard deviation of 4.07%.**

Keywords—THU-EP, EEG, Emotion recognition, Weakly supervised learning, Two-phase Multitask Autoencoder

I. Introduction

Emotions play a crucial role in human daily life as they directly influence judgment, memory, behavior, and social interactions [1]. Consequently, research on measuring human states, such as emotions and cognition, has been consistently conducted in the field of Human-Computer Interface (HCI) [2]. Since emotions manifest in the brain, analyzing Electroencephalogram (EEG) signals that measure this activity is essential for accurate emotion classification [3].

Recently, the development of emotion recognition models based on EEG signals has been actively pursued. In particular, EEG signal classification using deep learning techniques has shown promising results in this field [4]. Despite these advancements, the field faces significant challenges that hinder the practical application and commercialization of EEG-based emotion recognition systems.

One of the primary challenges is the high dependency on individual characteristics, as different individuals may exhibit varying EEG responses to the same emotional stimuli. This inter-subject variability poses a substantial obstacle in developing generalized models, which are essential for creating robust and scalable emotion recognition systems. Another challenge stems from the nature of the data collected during emotion-inducing activities, such as watching videos. Traditional methods often assign a single emotion label to the entire EEG recording from a video, even though emotional responses can fluctuate throughout the viewing experience. This approach can introduce label noise, reducing the accuracy and reliability of emotion classification models.

To address these challenges, this study proposes a novel deep learning model, Two-phase Weakly supervised Emotion Network (TWEN), designed specifically to enhance the generalization and robustness of EEG-based emotion recognition. To mitigate inter-subject variability, the TWEN model incorporates a two-phase multitask learning approach based on an AutoEncoder, which aims to reduce the Mean Squared Error (MSE) in both class-specific data and reconstruction data. This method facilitates the development of a more generalized emotion recognition model.

To address label noise, the proposed model introduces a weakly supervised learning framework with a Top-k Selection method. This approach extracts the k time frames with the strongest emotional responses for each class, ensuring that only the most relevant EEG data segments are used for training. By selectively retaining the most confidently predicted emotion labels, this method enhances the model's robustness to label noise and improves classification accuracy.

The structure of this paper is as follows: Section II provides a detailed description of the proposed TWEN model and its components. Section III presents the experimental methods, dataset, and results. Finally, Section IV discusses the conclusions of this study.

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) of the Korean government under Grants RS-2023-00208492, RS-2024-00422599, and 2020M3C1B8081320.

979-8-3503-6463-7/24/$31.00 ©2024 IEEE

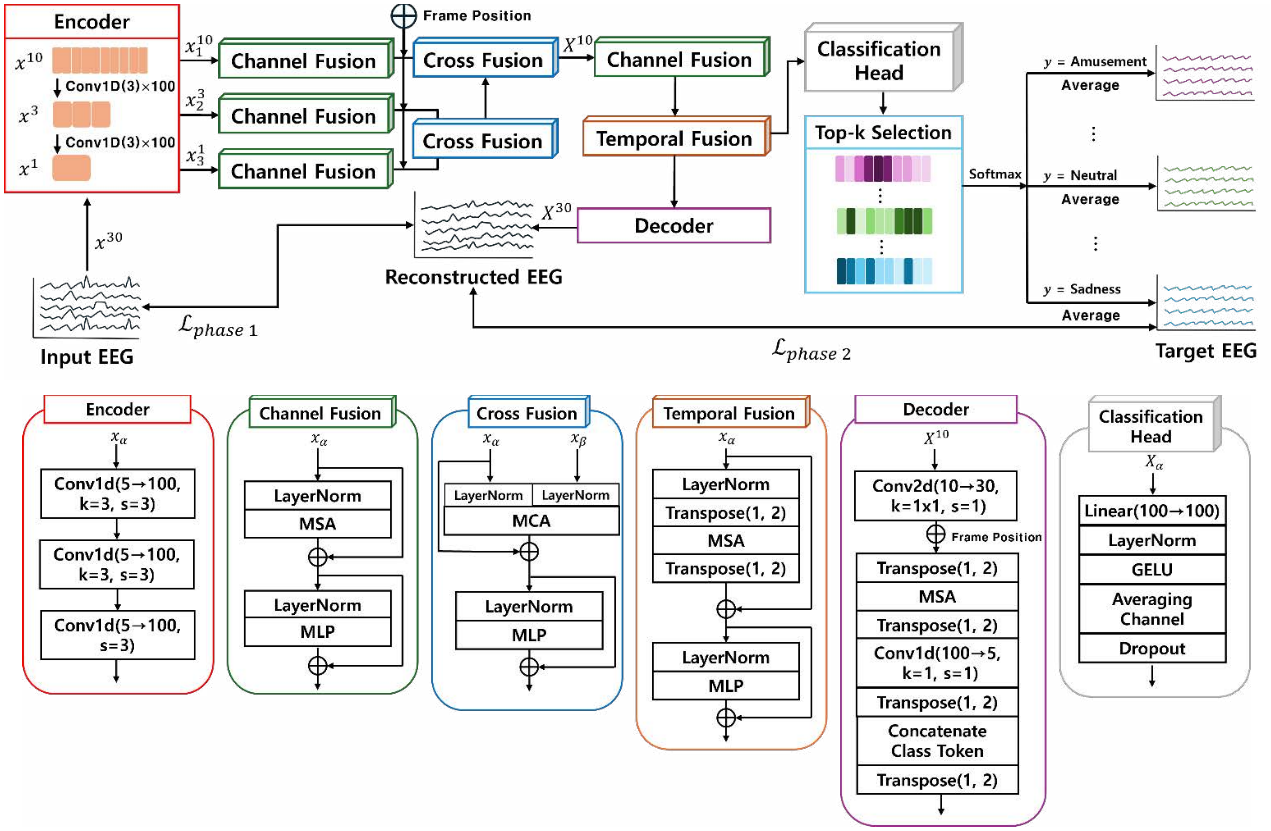

Fig. 1. The Overall TWEN Model Structure

II. Method

A. Model

The proposed model consists of an Encoder, three types of Fusion blocks, and a Decoder. It additionally employs Top-k Selection and a Two-phase approach. The overall model structure is presented in Fig. 1.

1) Encoder: The Encoder is designed to extract emotion-relevant EEG signals across multiple temporal scales by utilizing three sequential 1D convolutional layers (Conv1d) with a kernel size of 3 and a stride of 3. These layers progressively downsample the input 30-second EEG signals into temporal frames of 10, 3, and 1 second(s), capturing features at increasingly broader temporal scales, from local to global representations.

To incorporate positional information within these temporal frames, the Encoder employs sinusoidal encoding [5] to generate position embeddings. After performing channel fusion, these embeddings are added to the corresponding temporal frames, ensuring the model captures positional context at each scale. Additionally, average pooling with a kernel size of 3 and a stride of 3 is applied to create frame position embeddings, which are integrated into the different temporal frames, effectively merging temporal information with positional context.

2) Fusion:

Multi-Head Self Attention (MSA):

Multi-Head Self Attention (MSA) learns relationships within the same input data by generating Query, Key, and Value vectors for each element. Attention scores, calculated by comparing Query and Key vectors, weight the Value vectors to emphasize important patterns [5]. In this study, this process is performed across 4 attention heads to capture diverse aspects of the input data.

Multi-Head Cross Attention (MCA):

Multi-Head Cross Attention (MCA) learns relationships between two different input datasets by generating Query vectors from one and Key and Value vectors from the other. Attention scores highlight relevant information from the second input in relation to the first, with the weighted Value vectors creating a fused representation [5]. MCA also uses 4 attention heads to explore multiple relationships between the inputs.

a) Channel Fusion: Channel Fusion consists of Multi-Head Self Attention (MSA) and a Multi-Layer Perceptron (MLP). It is designed to fuse and extract channel features within the same input data.

b) Cross Fusion: Cross Fusion consists of Multi-Head Cross Attention (MCA) and a Multi-Layer Perceptron (MLP). It is designed to learn the relationships between two input datasets with different temporal lengths and integrate them into a single input dataset with a longer temporal frame, effectively combining local and global temporal features.

c) Temporal Fusion: Temporal Fusion consists of Multi-Head Self Attention (MSA) and a Multi-Layer Perceptron (MLP). It is designed to fuse and extract temporal features within the same input data.

3) Classification Head: The purpose of the Classification Head is to produce frame-level emotion scores. It achieves this by averaging the channels of the data that has been processed by Temporal Fusion, utilizing the GELU activation function. This channel-averaged data is then passed through a linear layer to generate 9 distinct emotion scores, which are subsequently classified into one of 9 emotion classes using a softmax function.

4) Decoder: The Decoder restores the fused data to the original input data dimensions using Conv2D, where both the kernel and stride sizes are set to 1. Positional information is incorporated via position embedding, and the data is reshaped to match the input format using Multi-Head Self Attention (MSA), Conv1D, and reshape layers. This sequence of operations produces the final reconstructed EEG data.

B. Implementation details

1) The Top-k selection: The Top-k selection method is applied to the frame-level emotion scores generated by the Classification Head. For each class, the top k frames with the highest scores are selected and averaged to produce the final emotion score for that class. This approach is designed to reduce label noise by focusing on the frames that exhibit the strongest emotional responses.

2) Two-phase Approach: The two-phase approach consists of the following steps:

a) First Phase: In the first phase, reconstructed data is generated through the decoder. Subsequently, the average of the reconstructed data is calculated to create class-specific target data for each emotion class. The loss function is designed to minimize the discrepancy between the subject's input data and the reconstructed data by combining Mean Squared Error ($\mathcal{L}_{MSE}$) and Cross-Entropy Loss($\mathcal{L}_{CE}$). This approach aims to capture generalized representations for each emotion class and minimize reconstruction errors. Equation (1) represents the loss function($\mathcal{L}_{p1}$)for the first phase.

$$\mathcal{L}_{p1} = \mathcal{L}_{CE}(y_1, label) + \mathcal{L}_{MSE}(recon, input) + \mathcal{L}_{MSE}(y_1, label) \qquad (1)$$

b) Second Phase: In the second phase, the model does not directly use the subject's data. Instead, it aims to minimize the discrepancy between the newly reconstructed data and the class-specific target data generated in the first phase. The loss function combines Cross-Entropy Loss($\mathcal{L}_{CE}$) and Mean Squared Error ($\mathcal{L}_{MSE}$) to reduce the mismatch between the reconstructed data and the class-specific target data. This ensures accurate emotion classification, aligns the reconstructed data more closely with the class-specific targets, enhances the model's generalization ability, and reduces subject dependency. Equation (2) represents the loss function($\mathcal{L}_{p2}$)for the second phase, with the weight of the Cross-Entropy Loss set to 0.001.

$$\mathcal{L}_{p2} = \lambda \times \mathcal{L}_{CE}(y_2, label) + \mathcal{L}_{MSE}(recon, input) + \mathcal{L}_{MSE}(y_2, label) \qquad (2)$$

III. Result

A. Dataset

In this study, we used the Tsinghua University Emotional Profiles (THU-EP) dataset for evaluation. The THU-EP dataset consists of data collected from 80 university students (50 females and 30 males, with a mean age of 20.16 years and an age range of 17-24 years). Each participant was exposed to emotion-inducing video clips across seven sessions. Each session included four video clips, making a total of 28 emotion-inducing video clips. Among these clips, 12 were intended to induce four negative emotions (Anger, Disgust, Fear, Sadness), 4 clips to induce a neutral emotion (Neutral), and the remaining 12 clips to induce four positive emotions (Amusement, Inspiration, Joy, Tenderness). To prevent emotional interference between blocks, participants solved 20 math problems between each session. The average length of the video clips was 67 seconds (ranging from 34 to 129 seconds) [6].

After viewing the video clips, participants reported their induced emotions on a scale of 0-7 for 12 emotions (Anger, Disgust, Fear, Sadness, Amusement, Joy, Inspiration, Tenderness, Arousal, Valence, Familiarity, Liking). EEG signals were recorded using the NeuSen.W32 wireless EEG system with 32 channels, with electrodes placed according to the international 10-20 system. The sampling frequency was 250Hz, and electrode impedance was kept below 10kOhm during the experiment.

B. Data preprocessing

Although EEG signals were recorded with 32 channels, only 30 effective channels were used, excluding the reference channels A1 and A2. To extract features from the raw EEG data, the signals were divided into 1-second intervals and a Short-Time Fourier Transform (STFT) was performed to compute the energy of five frequency bands (δ: 0.5-3Hz, θ: 4-7Hz, α: 8-13Hz, β: 14-29Hz, γ: 30-47Hz). The logarithm of these energies was then taken to calculate the Differential Entropy (DE). For this study, we used the last 30 seconds of each video clip to ensure consistent and intense emotion induction. Additionally, the original 12 emotion labels provided were re-labeled into 9 emotions to better match the emotions induced by the 28 video clips.

Consequently, the input data size was structured as SxCxTxB (S: number of samples, C: number of channels, T: time, B: frequency bands) with a size of 2240x30x30x5. Here, the number of samples (2240) represents the data from 80 participants for 28 video clips. We used the 10-fold cross-validation method to validate the model across the 80 participants. In the 10-fold cross-validation method, the entire dataset is divided into 10 subsets, with one subset used as the test set while the remaining subsets form the training set to validate the model. The training set was normalized for each channel and frequency band across all samples, and this normalization was applied to the test set.

C. Emotion recognition performance

The training was performed using the CosineAnnealing scheduler, which applied varying learning rates from a maximum of 3e-2 to a minimum of 1e-6. The number of epochs was set to 200, with a batch size of 64, and a weight decay of 0.05 was specified for the AdamW optimizer. The loss function used was Label Smoothing Loss with a smoothing factor of 0.1 for the 9 classes.

The Top-k selection method was applied, averaging the scores of the highest-scoring $\frac{10}{k}$ frames per class; for instance, when $k = 2$, this involved using 5 frame-level scores.

TABLE I. CLASSIFICATION ACCURACY(MEAN/STD) BY TOP-K AND PHASE ON THE THU-EP DATASET

Top-k	Phase	ACC(%)	STD(%)
1	1	56.43	3.34
	2	55.00	2.93
2	1	59.19	3.86
	2	**60.80**	4.07
5	1	59.02	1.96
	2	59.55	2.58
10	1	59.55	2.87
	2	58.97	3.65

Table I presents the classification accuracy (mean/standard deviation) for different values of Top-k and Phase on the THU-EP dataset. Notably, with a Top-k value of 2, the accuracy reaches its highest at 60.80%, with a standard deviation of 4.07% during Phase 2, highlighting the significance of these results. A performance improvement is observed when not all frame-level scores are used, compared to when Top-k is 1. The details are summarized in Table I.

TABLE II. PERFORMANCE COMPARISON ON THE THU-EP DATASET

Model	Acc(%)	STD(%)
CLISA [7]	45.7	11.8
MATCN [8]	58.6	5.7
TWEN (ours)	**60.8**	**4.07**

Table II presents a performance comparison of our proposed TWEN model against other emotion models on the THU-EP dataset. The results demonstrate that TWEN achieves the highest accuracy of 60.8% with a standard deviation of 4.07%, outperforming previous methods such as CLISA [7] and MATCN [8]. Specifically, TWEN surpasses CLISA by 15.1% in accuracy while also achieving a significantly lower standard deviation, indicating more consistent performance. Similarly, compared to MATCN, TWEN improves the accuracy by 2.2% with a lower standard deviation. These results establish TWEN as the new state-of-the-art for the THU-EP dataset.

Table III presents the accuracy (%) for each fold during Phase 1 and Phase 2 when k=2 on the THU-EP dataset. The results show that, overall, the accuracy in Phase 2 is either comparable to or higher than that in Phase 1, indicating an improvement in performance during the second phase.

IV. CONCLUSION

This study presents the TWEN, an innovative deep learning model that significantly advances EEG-based emotion recognition. TWEN incorporates several key strategies to overcome existing challenges in the field. First, it employs a weakly supervised learning framework with a Top-k Selection method to build a model that is robust against label noise. Second, by introducing a Two-phase Multitask Autoencoder, the model effectively reduces inter-subject variability, enabling the learning of more generalized emotion representations. Third, TWEN captures emotions occurring over both short and long-time frames by extracting local and global temporal features, which are then fused using Attention mechanisms to accurately recognize emotions across varying durations. These design and methodological innovations demonstrate TWEN's potential for broader application and commercialization in HCI and related fields,

TABLE III. ACCURACY(%) FOR EACH FOLD BY PHASE WHEN K=2 ON THE THU-EP DATASET

Fold	1	2	3	4	5	6	7	8	9	10	Average
Phase 1 Acc(%)	60.71	68.30	55.80	**54.46**	57.14	57.59	57.14	**61.16**	62.50	57.14	59.19
Phase 2 Acc(%)	**61.16**	**70.09**	**62.50**	54.02	**60.71**	**58.04**	**61.61**	57.14	**63.39**	**59.38**	**60.80**

particularly in areas where inter-subject variability poses significant challenges. Notably, when tested on the THU-EP dataset using the same validation methods as existing state-of-the-art models, TWEN surpassed the accuracy of those benchmark models. Using only EEG signals, the classification accuracy for nine emotions reaches 60.8%, which is insufficient for practical applications. However, by expanding the model to a multimodal emotion recognition system that integrates additional information such as facial expressions and vocal cues, it is expected that the accuracy can be significantly improved.

REFERENCES

[1] C. Mumenthaler, D. Sander, and A. S. R. Manstead, “Emotion recognition in simulated social interactions,” *IEEE Transactions on Affective Computing*, vol. 11, no. 2, pp. 308–312, 2018.

[2] S. Valenzi, L. Islam, F. Jurado, and M. D’Angelo, “Individual classification of emotions using EEG,” *Journal of Biomedical Science and Engineering*, vol. 7, no. 8, pp. 600-605, 2014.

[3] N. Jatupaiboon, S. Pan-Ngum, and P. Israsena, ″Real-Time EEG-Based Happiness Detection System,″ *The Scientific World Journal*, vol. 2013, pp. 618649, 2013.

[4] Y. Yang, Q. J. Wu, W.-L. Zheng, and B.-L. Lu, “EEG-based emotion recognition using hierarchical network with subnetwork nodes,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 408–419, 2017.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[6] W. Li, X. Hu, X. Long, L. Tang, J. Chen, F. Wang, and D. Zhang, “EEG responses to emotional videos can quantitatively predict big-five personality traits,” *Neurocomputing*, vol. 415, pp. 368–381, 2020.

[7] X. Shen, Z. Xie, J. Zhang, X. Hu, and S. Song, "Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition," *arXiv preprint arXiv:2109.09559*, 2021. [Online]. Available: https://arxiv.org/abs/2109.09559.

[8] X. Si, D. Huang, Y. Sun, and D. Ming, "Temporal Aware Mixed Attention-based Convolution and Transformer Network (MACTN) for EEG Emotion Recognition," *arXiv preprint arXiv:2305.18234*, 2023. [Online]. Available: https://arxiv.org/abs/2305.1823