2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669556

Emotion Transformer Fusion: Complementary Representation Properties of EEG and Eye Movements on Recognizing Anger and Surprise

Yiting Wang, Wei-Bang Jiang, Rui Li, Bao-Liang Lu*
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
wang.yiting@yandex.com, 935963004@sjtu.edu.cn, realee@sjtu.edu.cn, bllu@sjtu.edu.cn

***Abstract*—Emotion recognition plays an important role in diagnosing and treating many mental disorders as well as affective computing. Among six basic emotions, anger and surprise are relatively hard to be elicited in lab settings, and the complementary representation properties of encephalography (EEG) and eye movement signals on recognizing anger and surprise emotions remain unknown. Although the transformer architecture has the ability of parallelism which avoids many sequential operations as recurrent and convolutional layers, the knowledge of its performance and effectiveness on multimodal emotion recognition from EEG and eye movement signals is limited. To tackle these issues, we elaborately design the experiment and stimuli materials to effectively elicit surprise, anger, and neutral emotions, and propose an Emotion Transformer Fusion (ETF) model based on pure attention mechanism. Results of extensive experiments with multiple models on our dataset indicate that the complementary information of EEG and eye movements significantly improves the performance of discriminating anger, surprise and neutral emotions. Meanwhile, our proposed architecture outperforms baseline models with higher parallelism, which proves the capability of Transformer based architecture on multimodal emotion recognition with EEG and eye movement signals.**

***Index Terms*—multimodal emotion recognition, anger, surprise, transformer, attention**

I. Introduction

Multimodal emotion recognition draws raising attention in recent years, because it is the fundamental constitution of Brain Computer Interfaces (BCIs) and affective computing. Furthermore, not only do emotions become one of the useful tools evaluating mental health, their changing process is associated with some mental disorders like depression [1]. So, improving the ability of machine performing emotion recognition has great significance. Ekman's theory defines six basic emotions as happiness, fear, disgust, anger, surprise, and sadness [2]. There exists studies [3] investigating complementary representation of encephalography (EEG) and eye movement signals of six basic emotions except anger and surprise. In such manner, the complementary representation properties of EEG and eye movements for discriminating anger and surprise remain unknown.

Emotions are complicated psycho-physiological processes engaging with many internal and external activities which further complicated emotion recognition. Different modalities contain complementary information and reflect various aspects of emotions. Fusing modalities to take advantages of their attributes in terms of emotion recognition and construct robust models is promising [4]. Among many combinations of modalities, incorporate signals from external behaviors, e.g., eye movements, and the central nervous system, e.g., EEG, has been proven promising and efficient [3]. Considering the complexity nature of the neural mechanisms underlying the emotion processing, deep learning methods are widely applied in emotion recognition to automatically extract features. Because emotion transitions with temporal evolution are same among different modalities, models exploiting temporal features achieve encouraging performance on multimodal emotion recognition [3]. However, along the improvement of performance on the emotion recognition, networks became deeper with complex convolution and recursive structures. The hidden states in the network behave in a sequential manner resulting in a low degree of parallelization.

To tackle the problems mentioned above, we establish a new multimodal emotion dataset consisting of EEG and eye movement signals for three emotions: anger, surprise and neutral, and construct a multimodal network based on Transformer. To our best knowledge, although Transformer has become an influenced architecture in the natural language processing (NLP) and computer vision (CV) tasks, its applications to multimodal emotion recognition with EEG and eye movement signals remain limited. So, We propose the Emotion Transformer Fusion as a pure attention based model which combines Transformer encoders with attention based fusion, to utilize the parallelism and simplicity natures in emotion recognition with EEG and eye movement signals.

II. Experiment Setup

A. Experiment Detail

Seventeen Chinese subjects (9 males and 8 females) with ages of eighteen to thirty participated three sessions at different time. Each session lasted for around one hour with distinct content of stimuli. Eleven trials are designed in every session as shown in Fig. 1. Although our goal is investigating anger and surprise, we add neutral emotion in the experiment and final classification, which mediates two extreme emotions for subjects and promotes better induction. Each trial contains

* Corresponding author

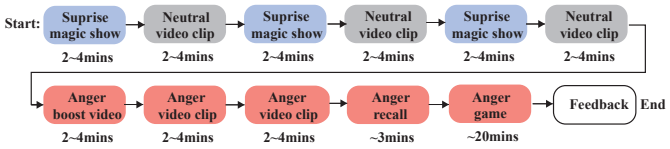

Fig. 1: The procedure of our experiment

two parts: emotion induction and self-assessment which allows subjects to score their emotion arousal level from one to ten.

The first six trials are designed to alternately trigger surprise and neutral emotions with video clips. Anger is hard to evoke for different subjects in lab settings and may have better effect with continuous stimulation. In this manner, five trials are arranged successively with the first one acting as boosting part, which is intended to transform subject's emotion state to angry quickly. The following two trials are video clips of social injustice. After three regular video trials, a special recall segment requires subjects to remember the recent event that made them most angry for around three minutes. At last, a delicately designed game is played by the subject for about twenty minutes. During the experiment, both EEG and eye movement signals are collected simultaneously except for the boosting trial. After the experiment, subjects are interviewed for pointing out the angriest period when they played the game and three trials to keep the data of each class and trial balanced.

B. Data Preprocessing

EEG signals consist of brain and non-brain contributions which brings difficulty of recognizing and analyzing brain-related EEG activities. So, we employ data preproccessing on the EEG signals to eliminate artifacts. Differential entropy (DE) feature is adopted which is one of the most effective EEG features in EEG based emotion recognition [5]. DE features are extracted in five frequency bands: δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, and γ: 31-50 Hz. Since 62 channels of EEG signals are collected, each sample has 310 (5 frequency bands multiple 62 channels) dimensional features.

As for the eye movement signals, we use PCA to remove light reflect to improve the quality of emotion information in the pupil diameter [4]. We extracted 23 features for eye movement signals whose detail is shown in Fig 2, where the number in the block stands for the dimension, and four features colored in blue are also event statistics.

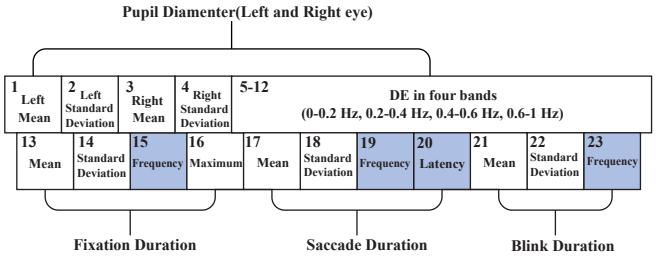

Fig. 2: Specific eye feature in every dimension.

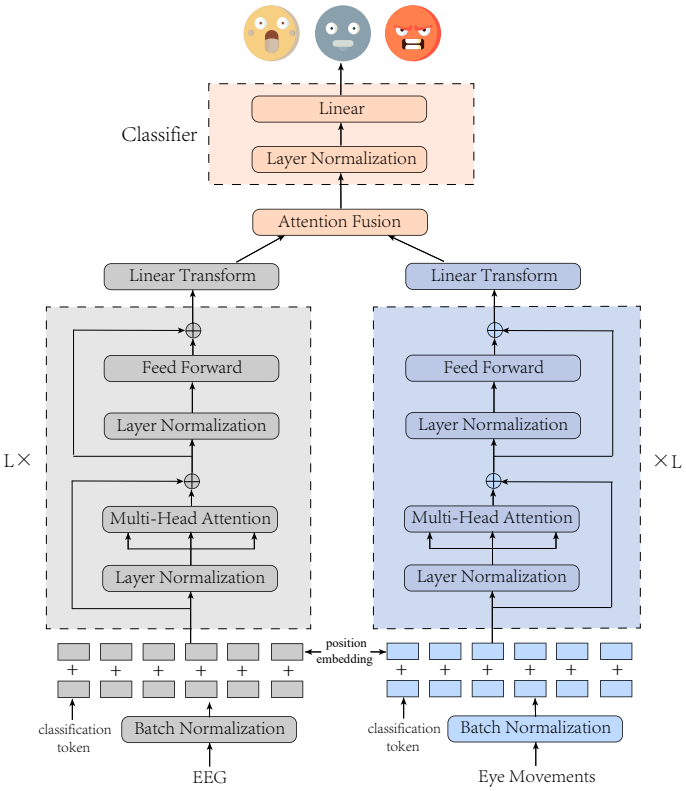

Fig. 3: The overall structure of our proposed ETF for multi-modal emotion recognition.

III. Methodology

A. Model Overview

We present the structure of our model in Fig. 3, where the left and right side in gray and blue are Transformer encoder for EEG and eye movements respectively. After position encoding, sequences from two modalities are fed into corresponding Transformer encoders and then fused to joint representation space by attention based fusion layer. Since we do not need a sequence as output, a fully connected feed-forward network as a classifier is performed instead of a decoder. Given the input EEG $X_{eeg} = (x^1_{eeg}, x^2_{eeg}, ..., x^T_{eeg}) \in R^{B\times T\times D_{eeg}}$ and eye movement signals $X_{eye} = (x^1_{eye}, x^2_{eye}, ..., x^T_{eye}) \in R^{B\times T\times D_{eye}}$, where B denotes batch size, T is the overlapping window size and D_{eeg}, D_{eye} are feature dimension of EEG and eye movements correspondingly, our model outputs a vector $O = (o_1, o_2, o_3)$ where o_i denotes the probability that the emotion is recognized as class i.

B. Signal Segments and Encoding

Since feature size of EEG and eye movement signals especially the latter is relatively small, we do not perform dropout in our model. To normalize data, alleviate over-fitting and increase learning speed, we apply a Batch Normalization [6] on mini-batch at the beginning. Considering the temporal sequences of EEG and eye movement are too long to be fed into the network directly, we deploy an overlapping windows

with the size of T seconds on the original signal, which will keep the total sample size (roughly 1200s per experiment) unchanged. We chose T as 5 seconds in our work.

An extra learnable classification token is concatenated at the beginning of the sequence to perform classification. Then position embeddings are added to the patches element-wise to obtain the information of time series. We deploy learnable 1D position embeddings for both EEG and eye movement signals in this paper.

C. Encoder

L identical layers are stacked to assemble the encoder, each of which consists of two sub-layers: a multi-head self-attention, and a fully connected feed forward network. Every sub-layer is started with layer normalization to relief internal covariate shift. And a residual connection is around each sub-layer to retain the information of the input feature and enhance the model stability.

1) Multi-Head Self-Attention: The attention function maps a query and a set of key-value pairs to an output, where the output is calculated as a weighted sum of the values. The weight assigned to each value is computed by query and corresponding keys. We employ the scaled dot product attention since the scale factor $\sqrt{d_k}$ avoids extremely small gradients after $softmax$. Specifically, dot product is performed on the query with all keys, which is divided by $\sqrt{d_k}$. Then a $softmax$ function is applied to obtain the weights for the values. We denote query $Q \in R^{t\times d_k}$, key $K \in R^{t\times d_k}$, value $V \in R^{t\times d_v}$ and output $O \in R^{t\times d_{model}}$ all as matrix, where t is the length of sequence, d_k is the dimension of query and key, d_v is the dimension of value and d_{model} is the output dimension of the encoder. The attention is computed as (1)

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \quad (1)$$

To obtain information from different representation subspaces of each modality at different positions, we combine several attention functions to achieve multi-head attention. The queries, keys and values are linearly projected h times with learned linear projects to d_k, d_k, d_v dimensions respectively, where h denotes number of heads. Then the attention functions are performed in parallel on projected queries, keys and values to calculate the $d_v - dimensional$ outputs. The outputs of all heads are concatenated and linearly projected to deliver $d_{model} - dimensional$ results to the next feed forward sub-layer. The calculation of multi-head attention is shown below:

$$MultiHead(Q,K,V) = Concat(O_{h_1},...,O_{h_h})W^O, \quad (2)$$

where $O_{h_i} = Attention(QW_i^Q, KW_i^K, VW_i^V)$ and learnable projection matrices $W_i^Q \in R^{d_{model}\times d_k}$, $W_i^K \in R^{d_{model}\times d_k}$, $W_i^V \in R^{d_{model}\times d_v}$ and $W^O \in R^{hd_v\times d_{model}}$.

2) Feed Forward Network: The second sub-layer is a simple fully connected feed forward network containing two linear projections with a Gaussian Error Linear Unit (GELU) activation in between. We use GELU rather than rectified linear unit (ReLu) because existing study [7] has shown that it has better performance and avoids vanishing gradients problem.

We employ the attention based fusion strategy since our model is designated to be based on the pure attention mechanism. The attention weight W_{init} is randomly initialized. Then we compute the inner product of attention weight with transformed features of two modalities, following a $softmax$ to normalize the results. After acquiring attention weights w_{eeg} and w_{eye}, the fused output O_{fuse} is calculated as weighted sum of single modality output. The whole process is formulized as following:

$$w_{eeg}^{'} =< O_{eeg}, W_{init} >, \quad (3)$$

$$w_{eye}^{'} =< O_{eye}, W_{init} >, \quad (4)$$

$$w_{eeg}, w_{eye} = softmax(w_{eeg}^{'}, w_{eye}^{'}), \quad (5)$$

$$O_{fuse} = w_{eeg}O_{eeg} + w_{eye}O_{eye}. \quad (6)$$

IV. Experiments

A. Implementation Details

1) Experimental Settings: We perform the three-fold-validation for each subject on our dataset. Performance of all models is evaluated by the averaged accuracy of three folds across all experiments. There are four hyper-parameters in our multimodal architecture: number of Transformer encoder layer L, dimension of fusion output O_{fuse}, learning rate and weight decay of Adam optimizer.

B. Results Analysis and Comparison

The baseline models of single modality are SVM, LSTM, and ET denotes Emotion Transformer which is a single Transformer of proposed model without attention based fusion. We practice LSTM as baseline model because LSTM and its variants are widely applied in the emotion recognition and its characteristics of utilizing temporal features makes it comparable with Transformer. As the result shown in the Table I, EEG significantly outperforms eye movement signals on recognizing three emotions of our dataset. Proposed Emotion Transformer exceeds baseline models with the accuracy of 83.3% and 77.86%, and standard deviation of 10.97% and

TABLE I: Average accuracy and standard deviations (acc/std %) of each single modality on different models

Model	EEG		Eye Movements	
	Acc.	Std.	Acc.	Std.
SVM	68.83	13.73	58.52	16.55
LSTM	79.77	**10.11**	73.32	**12.58**
ET	**83.30**	10.97	**77.86**	13.46

TABLE II: Average accuracy and standard deviations (%) of multimodal recognition using different models

Model	LSTM+AF	BDAE	ETConcat	ETF
Acc.	81.11	86.36	88.75	**90.02**
Std.	09.96	09.67	08.34	**07.06**

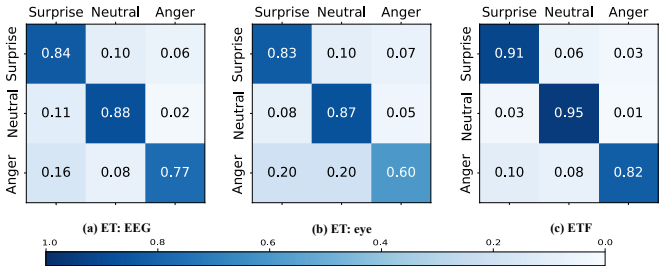

Fig. 4: Confusion matrices of different modalities and models.

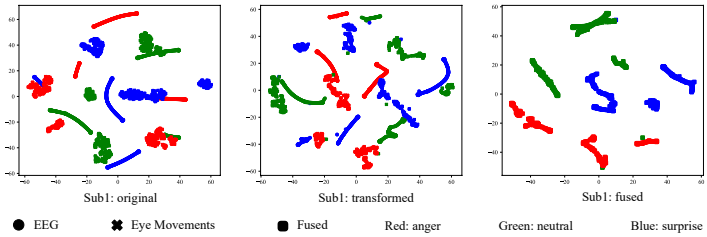

Fig. 5: Visualization of original features, transformed features and fused feature of ETF.

13.46% on EEG and eye movement signals respectively. Since we keep the transformer encoder as similar as possible with the original one proposed by Vaswani *et al.* [8], this shows the effectiveness of Transformer recognizing emotions.

For multimodal, all models are based on joint representations in order to reduce the variables for performance evaluation. Baseline models are bimodal deep auto-encoder (BDAE), LSTM with the same fusing strategy as proposed model, and ETConcat which replaces the attention based fusion with direct concatenation. As the result shown in the Table II, both multimodal LSTM and Transformer achieve better performance than those of each single modality, which proves that combining modalities brings considerable improvement on emotion recognition performance. In the meantime, our proposed model ETF achieves highest accuracy of 90.02% and lowest standard deviation of 7.06% suggesting the efficacy of our model. As the last two columns shown in Table II, accuracy and standard deviation of ETConcat is worse than those of ETF, which indicates that attention based fusion is able to extract more emotion related features from EEG and eye movement signals than the direct concatenation.

1) Confusion Matrices: Fig. 4 presents the confusion matrices of each single modality and multimodal of our proposed model, where each column represents the predicted emotion classified by models and each row serves as the target emotion class. The element (t, p) in the confusion matrix is the accuracy of samples in class t that was classified as class p. It is obvious that anger is harder to recognize compared with neutral and surprise. In our proposed model, ETF greatly improves the ability of recognizing anger. As for surprise, EEG and eye movements have similar abilities to recognize surprise emotion state, which is improved by ETF by around 8%. The observation indicates that EEG and eye movements contain complementary information on discriminating anger, surprise and neutral emotions. Moreover, our proposed model has the ability to extract more emotion related features with higher parallelism which benefits the emotion recognition.

2) Visualization: We visualize the feature distribution of one subject in Fig. 5, where green, blue, and red stands for neutral, surprise and anger correspondingly. As we can see, different emotions distribute arbitrarily in the original features and may overlap with one another. After the Transformer, same emotion state tends to gather together, while tangles and uncertainties still exist. Finally, the fused features are distinguishable and same emotion is nicely clustered. This visualization further demonstrates the effectiveness of ETF on emotion recognition.

V. Conclusions

This paper implemented multimodal experiments to successfully elicit anger, surprise, and neutral emotions with various types of stimuli, which also illustrates that EEG and eye movement signals are complementary to recognizing those three emotions. Furthermore, a pure attention mechanism based Emotion Transformer Fusion is utilized for multimodal emotion recognition. The best accuracy of 90.02% and standard deviation of 7.06% of proposed multi-modalities model have confirmed that Transformer based architecture with attention based fusion works efficiently on multimodal emotion recognition. The recognition results of single modality also suggest the efficacy of Transformer used in EEG or eye movements.

Acknowledgments

This work was supported in part by grants from the National Natural Science Foundation of China (Grant No. 61976135), SJTU Trans-Med Awards Research (WF540162605), the Fundamental Research Funds for the Central Universities, and the 111 Project.

References

[1] A. V. Bocharov, G. G. Knyazev, and A. N. Savostyanov, "Depression and implicit emotion processing: An EEG study," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 47, no. 3, pp. 225–230, 2017.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of Personality and Social Psychology*, vol. 17, no. 2, p. 124, 1971.

[3] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 176–183.

[4] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2011.

[5] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 2013, pp. 81–84.

[6] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[7] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.