2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385505

A Multitask Framework for Emotion Recognition Using EEG and Eye Movement Signals with Adversarial Training and Attention Mechanism

1st Wei Liu
Clinical Neuroscience Center
Ruijin Hospital
Shanghai Jiao Tong University School of Medicine
Shanghai, China
liuwei@shsmu.edu.cn

2nd Yun Luo
Clinical Neuroscience Center
Ruijin Hospital
Shanghai Jiao Tong University School of Medicine
Shanghai, China
ly41028@rjh.com.cn

3rd Yi Lu
International Department
Shanghai Luwan Senior High School
Shanghai, China
ly13524672061@163.com

4th *Yong Lu
Clinical Neuroscience Center
Ruijin Hospital
Shanghai Jiao Tong University School of Medicine
Shanghai, China
18917762053@163.com

***Abstract*—Affective brain-computer interface is one of the research frontiers which could promote the development of artificial intelligence and which might help the diagnosis and treatment of mental health diseases. In the field of emotion recognition with EEG and eye movement signals, however, it is challenging to build a model that can extract emotion-related information, fuse multiple modalities, and preserve the complementary characteristics at the same time. In this paper, we proposed a multi-task framework with adversarial training and attention mechanism (ATAM) for emotion recognition using EEG and eye movement signals. An adversarial training scheme is designed by maximizing mutual information loss within the same modalities and minimizing the cosine similarity loss between different modalities. With this design, the proposed ATAM not only preserved emotional information in EEG and eye movement signals, but also kept the complementary property between these two modalities. Attention mechanism was used to fuse multimodal features adaptively.Our evaluation of the model using the SEED and SEED-IV datasets revealed that the proposed method surpassed existing methodologies in recognition accuracy. Further analysis of the loss curves and attentional weight distributions indicated the effectiveness of ATAM in transforming and incorporating multi-modal properties. The adversarial training set-up, epitomized in our ablation study, was necessary for good performance.**

***Index Terms*—EEG, eye movement, emotion, adversarial training, attention mechanism**

I. Introduction

Emotions play a pivotal role in shaping human behavior, decision-making, and interpersonal interactions. As our world becomes increasingly digital and interconnected, it is crucial to build reliable, accurate, and objective emotion recognition models. Deep learning has accelerated the development of artificial intelligence [1]. Advanced models in computer vision [2], natural language processing [3], and speech recognition [6] have achieved significant results, and these methods have also been applied to physics [4], chemistry [5] and many other fields. However, researchers still face many obstacles in the affective computing field [7], [8]: 1) non-physiological signals, such as facial expressions and texts, are controlled by people, and these signals are not suitable for building objective emotion recognition models. 2) Physiological signals such as EEG and eye movement signals are appropriate for emotion recognition, but there is no standard deep learning framework for processing these signals.

One of the challenges in building emotion recognition model is the fusion of EEG and eye movement signals while preserving the complementary information. Lu and colleagues adopted feature-level fusion and decision-level fusion strategies to build multimodal emotion recognition models, and they found that multimodal features could improve the performance of emotion recognition models and that EEG and eye movement features were complementary to each other [9]. Zhang and colleagues built a heterogeneous convolutional neural networks and multimodal factorized bilinear pooling, aiming to express the complex internal relationship among multiple modalities [10]. Chao, Cao and Liu proposed a residual graph attention neural network to learn the connection relationship between EEG channels [11]. Chen *et al.* proposed a multi-stage multimodal dynamical fusion network, allowing to exploit much more fine-grained unimodal, bimodal and trimodal intercorrelations [12]. Liu and colleagues used several methods to fuse multiple modalities, such as biomdal deep autoencoder (BDAE) [15], deep canonical correlation analysis (DCCA) [16], deep generalized CCA [17], attention-based

* corresponding author

fusion strategy [18], and they also analyzed these fusion strategies from different aspects.

However, previous study showed that CCA-based transformation could remove noise from EEG and eye movement signals and preserve shared emotional-related information [18], there is a potential disadvantage in CCA-based methods: the complementary information might be lost after CCA regulation because CCA regulation maximizes the correlation between transformed multimodal features, leading to similar distributions regardless of the modalities.

In this paper, we proposed a multi-task framework with adversarial training and attention mechanism (ATAM) for identifying different emotions. ATAM first uses mutual information loss to obtain modality-related emotion information, then uses cosine similarity loss across different modalities to control the transformation process so that different modalities are perpendicular to each other. The mutual information loss and cosine similarity loss together forms an adversarial training scheme because they have opposite optimization directions, and the adversarial training scheme makes ATAM preserve emotion-related information and keep the complementary characteristics at the same time. In addition, attention mechanism is adopted to fuse EEG and eye movement features adaptively, and the fused feature is used to classify different emotions.

The contributions of this paper are as follows:

- We proposed the ATAM model which uses a multi-task framework where both adversarial training strategy and attention mechanism based multimodal fusion are adopted.
- We used the mutual information loss to regulate the original features and the transformed features for each modality so that we could extract the most modality-specific information.
- We used the cosine similarity loss between different modalities to preserve the complementary information.
- We assessed the model's effectiveness by monitoring loss curves, visualizing the distribution of attentional weights, and conducting ablation studies.

II. Method

In this section, we describe the proposed ATAM model in detail. Fig 1 demonstrates the structure of ATAM model.

A. Feature transformation

In the present study, we used EEG and eye movement data with a motive to recognize emotional patterns. Here, $X_1 \in \mathcal{R}^{N_1 \times d_1}$ represents the EEG feature matrix, while $X_2 \in \mathcal{R}^{N_2 \times d_2}$ delineates the eye movement feature matrix; N_1 and N_2 define the sample numbers, and d_1 and d_2 indicate the feature dimensions of EEG and eye movement features, respectively. Our experimental setup ensures EEG and eye movement signals are recorded simultaneously, hence $N_1 = N_2 = N$.

For the ATAM, the first stage incorporates non-linear transformation networks for EEG and eye movement features. In order to achieve this, we construct two deep neural networks $f_{eeg}(X_1; \theta_1)$ and $f_{eye}(X_2; \theta_2)$ where θ_1 and θ_2 are parameters for these two neural networks.

Following the processing, we subsequently obtained the transformed features related to EEG and eye movement:

$$O_1 = f_1(X_1) \tag{1}$$

$$O_2 = f_2(X_2) \tag{2}$$

B. Mutual information

The EEG signals and eye movement signals, recorded while the subjects were in different emotional states, contain a certain level of emotional information crucial for emotion recognition tasks. In the case of transformation functions f_1 and f_2, it is essential to preserve this emotional information in both EEG and eye movement signals. Thus, the transformed feature matrices O_1 and O_2 must maintain significant mutual information with the original features, denoted as X_{eeg} and X_{eye}, respectively.

According to [13], for random variables $\mathcal{X}$ and $\mathcal{Z}$, we can choose $\mathcal{F}$ to be the family of functions $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathcal{R}$, parameterized by a deep neural network with parameters signified as Θ. So we exploit the bound:

$$I(X, Z) \leq I_\Theta(X, Z) \tag{3}$$

$$I_\Theta(X, Z) = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}]) \tag{4}$$

Herein, $I(X, Z)$ represents the actual mutual information between $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$, $I_\theta(X, Z)$ is the estimated mutual information by function $\mathcal{F}$, $\mathbb{P}_{XZ}$ is the joint distribution of $\mathcal{X}$ and $\mathcal{Z}$, and $\mathbb{P}_X \otimes \mathbb{P}_Z$ is the product of marginal distributions. The Eq (4) can be estimated using samples from $\mathbb{P}_{XZ}$, and $\mathbb{P}_X \otimes \mathbb{P}_Z$ or by shuffling the samples from the joint distribution along the batch axis.

In this paper, we built deep neural networks M_{eeg} and M_{eye} to estimate the mutual information between original EEG (eye movement) features and the transformed EEG (eye movement) features.

In this paper, the network structure for mutual information estimation is a multi-layer non-linear transformation network. The first linear layer transformed the input features to new features of 100 dimensions, and the output is processed by ReLU activation function. The second linear layer take the 100-dimensional feature as input, and output a 50-dimensional features, and the ReLU processed output of the second layer is then fed into the third layer and is transformed into a scalar which is the estimated mutual information. The code for this model can be found by this link: https://github.com/csliuwei/ATAM/issues/1#issue-1989513174.

We calculated the estimated mutual information by shuffling the EEG and eye movement samples. Since the optimization process of deep learning seeks to minimize the value, we set the mutual information loss as the negative of the estimated mutual information:

$$L_{MI-eeg} = -1 \times M_{eeg}(X_1, O_1) \tag{5}$$

$$L_{MI-eye} = -1 \times M_{eye}(X_2, O_2) \tag{6}$$

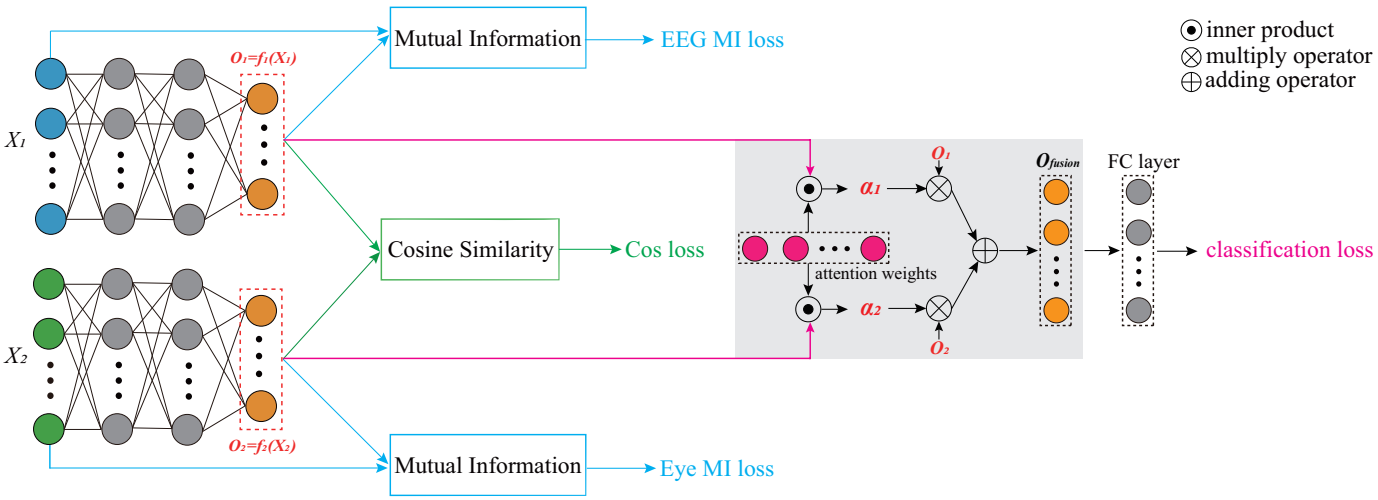

Fig. 1. The structure of ATAM.

C. Complementary information

According to [9], EEG and eye movement modalities exhibit complementary characteristics. Establishing a reliable emotion recognition model necessitates proficient exploitation of complementary information across different modalities. While we have preserved emotion-related information through mutual information regulation, it remains undetermined if the transformed EEG and eye movement features still retain their complementary attributes.

In this paper, cosine similarity was utilized to safeguard these complementary traits. For the transformed features O_1 and O_2, the cosine similarity was calculated within them:

$$L_{cos} = \cos(O_1, O_2) \tag{7}$$

As cosine similarity measures the cosine distance between two vectors, a higher similarity between vectors results in a larger cosine similarity. By minimizing the cosine distance, the transformed EEG and eye movement features achieved orthogonality. Subsequently, we considered EEG and eye movement features to be independent, possibly preserving their complementary information.

It is notable that we simultaneously maxmized the mutual information (i.e. minimizing the negative of mutual information) for each modality and minimized the cosine distance between different modalities. These two optimization tasks possess divergent optimization objectives, thus establishing an adversarial training scheme.

With this adversarial training framework, ATAM was able to extract emotional information from EEG and eye movement modalities whilst preserving their complementary properties.

D. Multimodal fusion

Another challenge for multimodal deep learning is to find a proper way to fuse multiple modalities. In this paper, we implemented an attention-based fusion strategy, and the process was depicted in Fig 1 with grey background.

For the attention-based fusion module (signified by the grey background layers in Fig 1), we first initialized an attention layer with parameters W_{attn}. Following this, we calculated the inner product of attentional weights and outputs of different modalities. The results were then normalized through the application of a softmax function to obtain attentional weights α_1 and α_2, respectively.

$$\hat{\alpha}_1 = < O_1, W_{attn} >, \tag{8}$$

$$\hat{\alpha}_2 = < O_2, W_{attn} >, \tag{9}$$

$$\alpha_1, \alpha_2 = softmax(\hat{\alpha}_1, \hat{\alpha}_2), \tag{10}$$

where W_{attn} was the hyperparameter to calculate attentional weights. After calculating the attentional weights, we extracted fused features as follows:

$$O = \alpha_1 O_1 + \alpha_2 O_2 \tag{11}$$

This attention-based fusion strategy can be conceptualized as an adaptive weighted sum, attributing weights to EEG and eye movement features relative to the input features and enabling analysis of the EEG and eye movement modalities contribution.

E. Classification

Subsequently, we added a fully-connected (FC) layer as a classifier to calculate the cross-entropy for classification loss. With all the loss functions we had built, the final loss function for optimization is as follows:

$$L = \gamma_1 L_{classification} + \gamma_2 (L_{MI-EEG} + L_{MI-Eye}) + \gamma_3 L_{cos} \tag{12}$$

where L was the total loss, and γ_1, γ_2, and γ_3 were hyperparameters. Finally, the back-propagation algorithm was used to update the weights of ATAM.

III. Experiment and Results

A. Dataset

1) SEED dataset: The SEED dataset was developed by Zheng and Lu [14]. Fifteen Chinese film clips of three emotions (happy, neutral and sad) were used as stimuli in the experiments. Every participant took part in the experiment three times. In this paper, we used the dataset as in our previous work [9], [15], [19] for the comparison study (9 participants 27 sessions because only 9 subjects had both EEG and eye movement signals). The SEED dataset contains EEG signals and eye movement signals.

2) SEED-IV dataset: The SEED-IV dataset was first used in [20]. Seventy-two film clips were chosen as stimuli materials. The dataset contains emotional EEG signals and eye movement signals of four different emotions, *i.e.*, happy, sad, neutral, and fear. Fifteen subjects (7 male and 8 female) participated in the experiments for three sessions, and the three sessions were performed on different days.

B. Feature extraction

1) EEG feature extraction: For EEG signals, we extracted differential entropy (DE) features using short-term Fourier transforms with a 4-second Hanning window without overlapping [21], [22]. Shi and colleagues [22] first proposed the DE feature, and proved that EEG signals within a short time period in different frequency bands are subject to a Gaussian distribution by the Kolmogorov-Smirnov test, and the DE features can be calculated by Eq. (13).

$$h(\mathbf{X}) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right) dx = \frac{1}{2}\log 2\pi e\sigma^2. \tag{13}$$

We extracted DE features from EEG signals in five frequency bands for all channels: delta (1-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-31 Hz), and gamma (31-50 Hz). There were in total $62 \times 5 = 310$ dimensions for 62 EEG channels. The linear dynamic system method was used to filter out noise and artifacts [23].

2) Eye movement features: The eye movement features extracted from SMI ETG eye-tracking glasses[1] contained both statistical features and computational features. Table I shows all 33 eye movement features used in this paper.

C. Model training

For the SEED dataset, the DE features of the first 9 movie clips were used as training data, and those of the remaining 6 movie clips were used as test data. In this paper, we built 'session-dependent' models for three emotions (happy, sad, and neutral), which was the same as in previous work [9], [15], [19].

For SEED-IV dataset, we used the data from the first 16 trials as the training data and the data from the remaining 8 trials as the test data [20]. ATAM was trained under 'session-dependent' setting to recognize four emotions (happy, sad, fear, and neutral)

For EEG and eye movement non-linear transform modules, we built a three layer neural network. For the first layer, the input dimension for EEG and eye movement modalities were 310 and 33, respectively, and the output dimensions for both modalities were randomly choose from $[100, 200]$. For the second layer, the output dimensions for both modalities were also randomly chosen from $[20, 50]$. For the third layer, the output dimension for both modalities was set to 12 according to previous results [18]. During training, the epoch number was 200, the batch size was 30, and the learning rate was 0.001. We randomly chose 10 groups of hyper-parameter (i.e., we randomly sampled 10 times for the three hidden layers of non-linear transform modules), and ran the model 10 times to find the best hyperparameters. For hyper-parameters in Eq. (12), we set $\gamma_1 = 0.5$, $\gamma_2 = 0.2$, and $\gamma_3 = 0.02$.

TABLE I
Summary of extracted eye movement features.

Eye movement parameters	Extracted features
Pupil diameter (X and Y)	Mean, standard deviation, DE in four bands (0–0.2Hz,0.2–0.4Hz, 0.4–0.6Hz,0.6–1Hz)
Disperson (X and Y)	Mean, standard deviation
Fixation duration (ms)	Mean, standard deviation
Blink duration (ms)	Mean, standard deviation
Saccade	Mean and standard deviation of saccade duration(ms) and saccade amplitude(°)
Event statistics	Blink frequency, fixation frequency, fixation duration maximum, fixation dispersion total, fixation dispersion maximum, saccade frequency, saccade duration average, saccade amplitude average, saccade latency average.

D. Experimental results

1) Loss curves: We first checked the adversarial training process. As we have described, the mutual information should be maximized so that the emotional information in both EEG and eye movement modalities could be preserved. In addition, since we also wanted to keep the complementaty property of EEG and eye movement signals, we minimized the cosine similarity between these two modalities so that the transformed features from different modalities were independent with each other.

For the SEED training set (panel a) in Fig 2, the mutual information for EEG modality and eye movement modality grew gradually, and the cosine similarity between the two modalities became smaller. Curves in Fig 2a had the same trends as expected: emotional information was preserved during the training process, and the complementary information was also kept.

[1]https://en.wikipedia.org/wiki/SensoMotoric_Instruments

For the SEED test set (panel b) in Fig 2, the curves had similar trends compared with panel a. This phenomenon indicates that the ATAM could learn some deeper relationship between these two modalities, and the learnt model was able to extract the emotional information while keeping the complementary information at the same time.

It was worth noting that the values of mutual information for eye movement signal of the SEED test set was negative, even though the curve had the correct trending. The reason for the negative mutual information might be that our mutual information estimator was a deep neural network, and the output layer was a linear transformation layer without any other regulations so the estimated value of mutual information could be any number. With more data collected in the future and carefully designed mutual information estimation network, the mutual information estimator for eye movement modality might have a better generalization performance.

The curves of the SEED-IV training set and test set were presented in Figs 3a and b. Similaly, the mutual information for both EEG and eye movement increased on both training set and test set during training, and the cosine similarity decreased on both training set and test set. These curves changed as expected, indicating that the ATAM learnt how to transform EEG and eye movement features effectively.

2) Emotion recognition performance: In this section, we compared the emotion recognition accuracies of ATAM and other models on SEED and SEED-IV datasets, and the results were listed in Table II and Table III, respectively.

For the SEED dataset, the ATAM model achieved 94.8% recognition accuracy in classfying happy, sad, and neutral emotions, and the proposed model performed best among 13 methods shown in Table II. However, the ATAM did not have the best standard deviation, which implied that the model might have performance fluctuations for different subjects.

The spatial-temporal recurrent neural network (STRNN) [27] and the region to global spatial-temporal neural network (R2G-STNN) [30] were models that learnt information from both spatial and temporal dimensions of EEG signals, and achieve 89.5% and 93.3% recognition accuracies. Both bi-hemispheres domain adversarial neural network (BiDANN) [28] and bi-hemispheres discrepancy model (BiHDM) [29] took the right and left hemispheres into consideration, and the recognition accuracies were 92.5% and 93.1%, respectively.

The regularized graph neural network (RGNN) [31] was a graph neural network model for EEG signals, and it had a recognition accuracy of 94.2%. MAE, short for masked autoencoder, was first proposed in [32]. Li and colleagues applied the MAE model to EEG-based emotion recognition tasks and the accuracy was 92.3% [26]. In [26], the authors also proposed a multi-view spectral-spatial-temporal masked autoencoder (MV-SSTMA) model, and the model achieved an accuracy of 95.3%. However, since the model training and evaluation process was different from the process in this paper, we did not compare with the MV-SSTMA.

TABLE II
THE COMPARISON OF MEAN ACCURACY RATES (%) AND STANDARD DEVIATIONS (%) OF VARIOUS METHODS ON THE SEED DATASET.

Methods	Mean (%)	Std (%)
STRNN [27]	89.5	7.6
BiDANN [28]	92.4	7.0
BiHDM [29]	93.1	6.1
R2G-STNN [30]	93.3	6.0
RGNN [31]	94.2	6.0
MAE [26]	92.3	5.2
Concatenation [9]	83.7	–
MAX [9]	81.7	–
Fuzzy Integral [9]	87.6	19.9
SLFN with subnetwork nodes [25]	91.5	–
Bimodal-LSTM [19]	94.0	7.0
BDAE [15]	91.0	8.9
ATAM	**94.8**	7.5

TABLE III
THE COMPARISON OF MEAN ACCURACY RATES AND STANDARD DEVIATIONS OF VARIOUS METHODS ON THE SEED-IV DATASET.

Methods	Mean (%)	Std (%)
DGCNN [24]	69.9	16.3
BiDANN [28]	70.3	12.6
BiHDM [29]	74.4	14.1
RGNN [31]	79.4	10.5
MAE [26]	87.8	5.0
Concatenation [18]	77.6	16.4
MAX [18]	60.0	17.1
Fuzzy Integral [18]	73.6	16.7
BDAE [20]	85.1	11.8
ATAM	**91.6**	10.0

Concatenation, MAX, and fuzzy integral were three multimodal fusion methods: concatenation was a feature-level fusion which means we concated the EEG and eye movement features together; MAX was a decision-level fusion which used the biggest decision value from different classifiers as the final decision value; And the fuzzy integral fusion method applied fuzzy integral to fuse probabilities from different classifiers. These three fusion strategies achieved 83.7%, 81.7%, and 83.7% accuracies, respectively.

The subnetwork model [25], bimodal-LSTM model [19] and bimodal deep autoencoder (BDAE) [15] were also previously proposed deep learning models, and they obtained 91.5%, 94.0%, and 91.0% accuracies, respectively.

For the SEED-IV dataset, as shown in Table III, the proposed ATAM also had the best 91.6% recognition accuracy on fear, happy, sad, and neutral emotion classification task among all 10 methods. Similar to results in Table II, the ATAM model did not have the best standard deviation, and we also did not compare the ATAM with MV-SSTMA model because of different training and evaluation process.

From the perspective of emotion recognition accuracies, the proposed ATAM could achieve best performance compared with many other models. However, since the standard deviations were not the best, the proposed ATAM model could be improved, and this will be one of our future work.

3) Attentional weights visualization: The attention-based fusion could be seen as an adaptive weighted sum, and

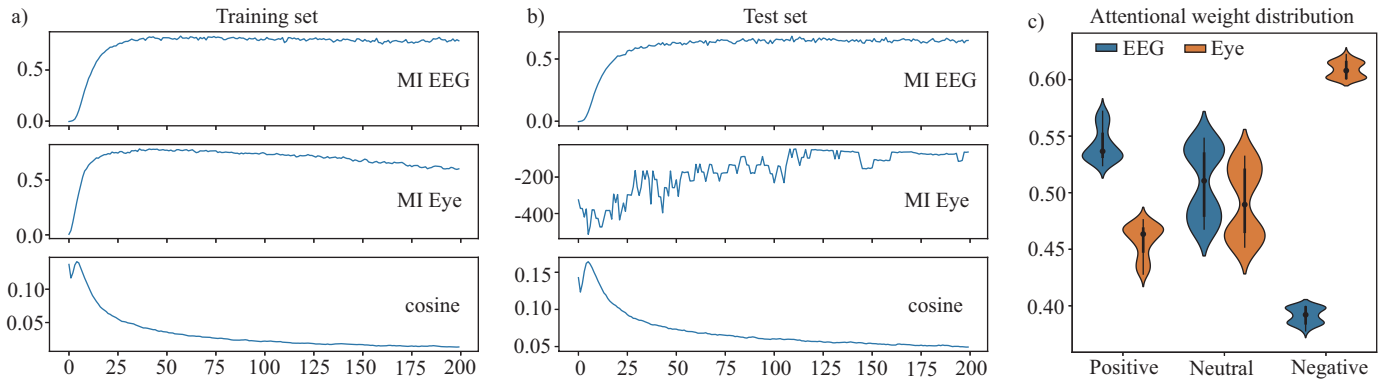

Fig. 2. The training process and attentional weight distributions on the SEED dataset. Panel a depicts the average training curves for mutual information of EEG signal, eye movement signal, and the cosine similarity. Panel b shows the average curves on test set, which means that we calculated these values on test set for every epoch. For panels a and b, the x-axis stands for the epoch, and the y-axis is the values for correponding metrics.'MI' means mutual information, and 'cosine' is cosine similarity. Panel c presents the attentional weight distribution of test set for different emotions.

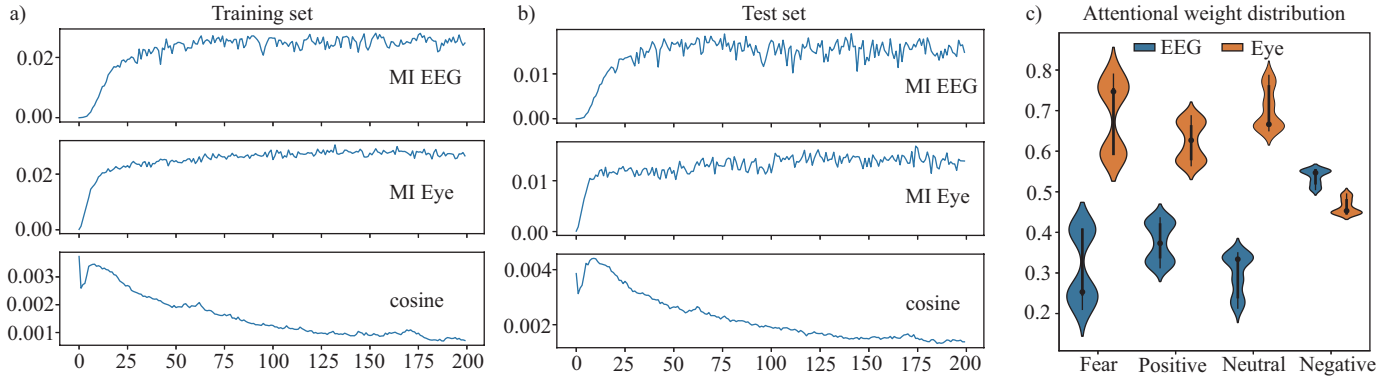

Fig. 3. The training process and attentional weight distributions on the SEED-IV dataset. Panel a depicts the average training curves for mutual information of EEG signal, eye movement signal, and the cosine similarity. Panel b shows the average curves on test set, which means that we calculated these values on test set for every epoch. For panels a and b, the x-axis stands for the epoch, and the y-axis is the values for correponding metrics. 'MI' means mutual information, and 'cosine' is cosine similarity. Panel c presents the attentional weight distribution of test set for different emotions.

by analysing the weights, we could know which modality contributed more for different emotions. The attentional weight distributions of test set of the SEED and SEED-IV dataset were depicted in Fig 2c and Fig 3c, respectively.

From Fig 2c, it was obvious that EEG features contributed more than eye movement features for positive emotion, eye movement features were more important for negative emotion recognition, and EEG features and eye movement features had almost the same contribution for neutral emotion. Lu and colleagues analyzed the complementary property between EEG and eye movement signals from the aspect of unimodal classification results [9]. Compared with the previous results in [9], the ATAM model revealed a similar conclusion: EEG and eye movement signals have complementary information on emotion recognition tasks, the ATAM model could preserve this complementary property, and the EEG features were better at positive emotion recognition while the eye movement features were better at negative emotions.

From Fig 3c, eye movement features had better performance than EEG features on fear, positive, and neutral emotions. EEG only obtained higher weights for negative emotion. Two things were worth noting: 1) the distributions of the SEED-IV dataset were different from the distributions of the SEED dataset. In the ideal situation, attentional weight distributions should be the same for the same emotion. However, because of limited data and individual differences in EEG signals, the proposed ATAM model returned different distributions. Furthermore, the fear emotion in SEED-IV dataset might have influences on the perception of other emotions, resulting in a different distribution. 2) The contributions of EEG and eye movement features for different emotions were different from results in [20]. The difference might be caused by different measurement: Zheng and colleagues [20] analyzed the contribution based on the unimodal classification results, while the ATAM calculated the contribution based on the transformed EEG and eye movement features in a multimodal aspect. Since different methods were used, it was reasonable that different modality contributions were found.

4) Ablation study: Finally, we evaluated the ATAM by removing mutual information loss or cosine similarity loss, and the results for the SEED and SEED-IV dataset were listed in Tables IV and V, respectively.

TABLE IV
THE RECOGNITION ACCURACIES OF ABLATION STUDY ON THE SEED DATASET.

Methods	Mean (%)	Std (%)
ATAM-no-MI	93.5	7.9
ATAM-no-Cos	94.1	8.2
ATAM	94.8	7.5

TABLE V
THE RECOGNITION ACCURACIES OF ABLATION STUDY ON THE SEED-IV DATASET.

Methods	Mean (%)	Std (%)
ATAM-no-MI	89.8	13.3
ATAM-no-Cos	90.8	13.4
ATAM	91.6	10.0

For the SEED dataset, after removing the mutual information loss, the emotion recognition accuracy is 93.5%. After removing the cosine similarity loss, the emotion recognition accuracy is 94.1%.

For the SEED-IV dataset, when removing the mutual information loss, the emotion recognition accuracy is 89.8%. When removing the cosine similarity loss, the emotion recognition accuracy is 90.8%.

According to results in Tables IV and V, we drew a conclusion that the mutual information and cosine similarity loss could work together, and the adversarial training scheme was necessary for EEG and eye movement feature transformation.

IV. CONCLUSION

In this paper, we have proposed a multi-task model with adversarial training and attention mechanism (ATAM) for emotion recognition using EEG and eye movement signals. We have designed an adversarial training scheme by maximizing mutual information between the original EEG (eye movement) features and the transformed EEG (eye movement) features and minimizing the cosine similarity between transformed EEG and eye movement features. With this adversarial training setting, the proposed ATAM not only preserved emotional information from EEG and eye movement signals, but also preserves the complementary characteristics of these two modalities. We have evaluated the ATAM model on the SEED and SEED-IV datasets, and the experimental results have revealed that the proposed method achieved the best recognition accuracies on both datasets comparing with many other existing methods. The analysis of loss curves and attentional weight distributions has proved some of our assumptions that the ATAM could transform and fuse multimodal features effectively. The ablation study has provided extra evidence that the adversarial training scheme was necessary for good model performance.

However, the standard deviations of the ATAM model were not the best compared with other methods, suggesting that the proposed model might have fluctuating performance among different subjects. Furthermore, the attentional weight distributions of the same emotion across different datasets were not consistent. In the future, we will evaluate the ATAM on more datasets, and we would try to build a model that has stable performance across different subjects and datasets.

ACKNOWLEDGMENT

This work was supported in part by grants from Shanghai 2022 "Science and Technology Innovation Action Plan" Artificial Intelligence Technology Support Project (No. 22511106000), STI 2030-Major Projects+2022ZD0208500, the National Natural Science Foundation of China (No. 61976135), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), SJTU Global Strategic Partnership Fund, Shanghai Marine Equipment Foresight Technology Research Institute 2022 Fund (No. GC3270001/012), SJTU Global Strategic Partnership Fund (2021 SJTUHKUST), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

REFERENCES

[1] LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), pp.436-444.
[2] Voulodimos, A., Doulamis, N., Doulamis, A. and Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. Computational intelligence and neuroscience, 2018.
[3] Bharadiya, J., 2023. A Comprehensive Survey of Deep Learning Techniques Natural Language Processing. European Journal of Technology, 7(1), pp.58-66.
[4] Ma, P., Petridis, S. and Pantic, M., 2022. Visual speech recognition for multiple languages in the wild. Nature Machine Intelligence, 4(11), pp.930-939.
[5] Mater, A.C. and Coote, M.L., 2019. Deep learning in chemistry. Journal of chemical information and modeling, 59(6), pp.2545-2559.
[6] Ma, P., Petridis, S. and Pantic, M., 2022. Visual speech recognition for multiple languages in the wild. Nature Machine Intelligence, 4(11), pp.930-939.
[7] Picard, R.W., 2000. Affective computing. MIT press.
[8] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W. and Zhang, W., 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 83, pp.19-52.
[9] Lu, Yifei, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining Eye Movements and EEG to Enhance Emotion Recognition. In IJCAI, vol. 15, pp. 1170-1176. 2015.
[10] Zhang, Y., Cheng, C., Wang, S. and Xia, T., 2022. Emotion recognition using heterogeneous convolutional neural networks combined with multimodal factorized bilinear pooling. Biomedical Signal Processing and Control, 77, p.103877.
[11] Chao, H., Cao, Y. and Liu, Y., 2023. Multi-channel EEG emotion recognition through residual graph attention neural network. Frontiers in Neuroscience, 17.
[12] Chen, S., Tang, J., Zhu, L. and Kong, W., 2023. A multi-stage dynamical fusion network for multimodal emotion recognition. Cognitive Neurodynamics, 17(3), pp.671-680.
[13] Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, D. (2018, July). Mutual information neural estimation. In International conference on machine learning (pp. 531-540). PMLR.
[14] Zheng, W.L. and Lu, B.L., 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on autonomous mental development, 7(3), pp.162-175.
[15] Liu, W., Zheng, W.L. and Lu, B.L., 2016. Emotion recognition using multimodal deep learning. In Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23 (pp. 521-529). Springer International Publishing.

[16] Qiu, J.L., Liu, W. and Lu, B.L., 2018. Multi-view emotion recognition using deep canonical correlation analysis. In Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V 25 (pp. 221-231). Springer International Publishing.
[17] Lan, Y.T., Liu, W. and Lu, B.L., 2020, July. Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.
[18] Liu, W., Qiu, J.L., Zheng, W.L. and Lu, B.L., 2021. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 14(2), pp.715-729.
[19] Tang, H., Liu, W., Zheng, W.L. and Lu, B.L., 2017. Multimodal emotion recognition using deep neural networks. In Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24 (pp. 811-819). Springer International Publishing.
[20] Zheng, W.L., Liu, W., Lu, Y., Lu, B.L. and Cichocki, A., 2018. Emotionmeter: A multimodal framework for recognizing human emotions. IEEE transactions on cybernetics, 49(3), pp.1110-1122.
[21] Duan, R.N., Zhu, J.Y. and Lu, B.L., 2013, November. Differential entropy feature for EEG-based emotion classification. In 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 81-84). IEEE.
[22] Shi, L.C., Jiao, Y.Y. and Lu, B.L., 2013, July. Differential entropy feature for EEG-based vigilance estimation. In 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 6627-6630). IEEE.
[23] Shi, L.C. and Lu, B.L., 2010, August. Off-line and on-line vigilance estimation based on linear dynamical system and manifold learning. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (pp. 6587-6590). IEEE.
[24] Song, T., Zheng, W., Song, P. and Cui, Z., 2018. EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Transactions on Affective Computing, 11(3), pp.532-541.
[25] Yang, Y., Wu, Q.J., Zheng, W.L. and Lu, B.L., 2017. EEG-based emotion recognition using hierarchical network with subnetwork nodes. IEEE Transactions on Cognitive and Developmental Systems, 10(2), pp.408-419.
[26] Li, R., Wang, Y., Zheng, W.L. and Lu, B.L., 2022, October. A Multi-view Spectral-Spatial-Temporal Masked Autoencoder for Decoding Emotions with Self-supervised Learning. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 6-14).
[27] Zhang, T., Zheng, W., Cui, Z., Zong, Y. and Li, Y., 2018. Spatial–temporal recurrent neural network for emotion recognition. IEEE transactions on cybernetics, 49(3), pp.839-847.
[28] Li, Y., Zheng, W., Cui, Z., Zhang, T. and Zong, Y., 2018, July. A Novel Neural Network Model based on Cerebral Hemispheric Asymmetry for EEG Emotion Recognition. In IJCAI (pp. 1561-1567).
[29] Li, Y., Wang, L., Zheng, W., Zong, Y., Qi, L., Cui, Z., Zhang, T. and Song, T., 2020. A novel bi-hemispheric discrepancy model for EEG emotion recognition. IEEE Transactions on Cognitive and Developmental Systems, 13(2), pp.354-367.
[30] Li, Y., Zheng, W., Wang, L., Zong, Y. and Cui, Z., 2019. From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition. IEEE Transactions on Affective Computing, 13(2), pp.568-578.
[31] Zhong, P., Wang, D. and Miao, C., 2020. EEG-based emotion recognition using regularized graph neural networks. IEEE Transactions on Affective Computing, 13(3), pp.1290-1301.
[32] He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R., 2022. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).