2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)
Virtual Conference, May 4-6, 2021

Multimodal Emotion Recognition Using a Modified Dense Co-Attention Symmetric Network

Zhi-Wei Zhao, Wei Liu and Bao-Liang Lu* *Fellow, IEEE*

***Abstract*— In this paper, an attention-based network called modified Dense Co-Attention Symmetric Network is investigated to classify three human emotions using electroencephalogram (EEG), eye movement features (EYE), and raw eye movement image (EIG) data. The key idea of this model is to use a novel co-attention layer, which augments the weights of the critical feature channel and builds correlations among the different modalities. We stack the co-attention layer to form a hierarchy that strengthens the effect layer-by-layer to more precisely predict the emotion states. In an emotion recognition task with three modalities (EEG, EYE and EIG), we achieve 87.63% accuracy. To deal with the problem that EEG signals are difficult and expensive to collect in practical applications, we achieve 86.28% accuracy with two modalities (EYE and EIG) and propose a new classification paradigm that uses EIG, EYE, and randomly selected EEG data that we have on hand, obtaining 86.81% accuracy.**

I. INTRODUCTION

Emotions play an important role in human life. Emotion recognition is also important in the field of human-computer interaction (HCI). The introduction of affective factors for HCIs has rapidly developed as an interdisciplinary research field called affective computing.

The electroencephalogram (EEG) is one of most common used physiological signals and has proved to be a reliable and suitable tool for emotion recognition. Duan *et al.* extracted the differential entropy (DE) feature of EEG signals and found it effective in emotion recognition tasks [1]. Zheng and Lu further compared different channels and critical bands of EEG signals with a deep neural network [2].

Multimodal approaches have recently been pursued in the field of emotion recognition. Lu *et al.* enhanced emotion recognition by combining EEG and eye movement (EYE) data [3]. As some of the key procedures in multimodal deep learning, various fusion methods such as Bimodal Deep AutoEncoder (BDAE) have been implemented to improve the performance of affective models [4]. Lan *et al.* introduced attention mechanism among three modalities and investigated the role of different modalities for emotion recognition [5].

However, there is still room for improvement of the existing approaches from the following three aspects. First, most multimodal emotion recognition methods have been based on EEG and EYE data, while eye image (EIG) data have been seldom exploited. Guo *et al.* found that the EIG data also contain emotion-related information [6]. Second, while various fusion methods have been developed, the correlation among modalities has been rarely investigated. The attention mechanism has great potential in multimodal fusion. Lan *et al.* used the attention mechanism, but without interaction among modalities [5]. Third, the existing multimodal approaches need to use EEG signals. However, EEG signals are difficult and expensive to collect, representing the bottleneck for the use of emotion recognition systems in practical applications. Thus, an emotion recognition model that does not need to collect EEG signals has higher practical value.

In this paper, we investigate EIG data by feeding it into VGG19 [7], which is pretrained on ImageNet to extract higher level features. To capture intermodality interactions, we modify the Dense Co-Attention Symmetric Network (DCAN), which takes dense symmetric interactions among input modalities into consideration. To fully exploit this characteristic, we propose a new classification paradigm that only needs to collect EYE and EIG data and uses randomly selected EEG signals that we already have on hand. This emotion recognition mode has more practical value due to its greater convenience and lower cost. Moreover, even if the EEG data are completely absent, we can still achieve a relatively high accuracy by using our proposed model.

II. METHODS

A. Dense Co-Attention Symmetric Network

Dense Co-Attention Symmetric Network (DCAN) was originally proposed to solve the visual question answering (VQA) problem [8]. It presents an architecture that enables dense, bidirectional interactions between two modalities and contributes to boost the prediction accuracy of answers [8]. The model contains three main parts: feature extraction, stacked dense co-attention layers, and prediction (Fig. 1 (a)).

To take the advantage of DCAN to interact and build the connections among the modalities, we adopt this model. Since the emotion recognition task is different from the VQA problem, we make several modifications to make DCAN suitable to the emotion recognition task. Our modified DCAN retains the same three parts (Fig. 1(b)) and is described in detail in section 2.B. The main modifications are listed as follows.

1) If the channel numbers of the question features are different from image features, we use simple matrix reshape

Zhi-Wei Zhao, Wei Liu and Bao-Liang Lu are with the Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, the Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, the Brain Science and Technology Research Center, Qing Yuan Research Institute, Shanghai Jiao Tong University, 800 Dong Chuan Rd., Shanghai 200240, China, and the Center for Brain-Machine Interface and Neuromodulation, RuiJin Hospital, Shanghai Jiao Tong University School of Medicine, 197 Ruijin 2nd Rd., Shanghai 200020,China.

*Corresponding author (bllu@sjtu.edu.cn)

2021 10th International IEEE/EMBS Conference on Neural Engineering (NER) | 978-1-7281-4337-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/NER49283.2021.9441352

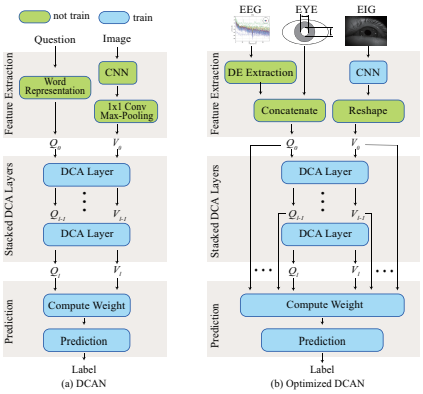


Fig. 1. Structure of the original dense co-attention symmetric network (a), and the modified DCAN proposed in this paper (b)

to retain original image features. However, the original DCAN uses 1×1 convolution kernel and max-pooling to adapt image features with question features.

2) We concatenate all of the outputs of the stacked DCA layers to strengthen the correlation while retaining the original features. The original DCAN only uses the last layer output of the stacked DCA layers.

3) In our emotion recognition tasks, we have three modalities, whereas the original DCAN can only merge two modalities. We first concatenate the EEG and EYE data into a single feature and then feed the concatenated features into our model with the raw EIG.

4) EIG features and concatenated features have different dimensions, but the original DCAN requires two features with the same dimension. We change the shape of a learnable weight matrix (W_i in Algorithm 1), so the features of the two modalities can have different dimensions.

5) We train a CNN while also extracting the image features because EIG data are different from the normal images targeted in object recognition problems. By contrast, in the original DCAN, the pretrained CNN for image feature extraction and the training part of the model are completely independent of each other.

B. Modified Dense Co-Attention Symmetric Network

1) Feature Extraction: Given the representations of EEG, EYE and EIG data, we extract features first.

For the EEG data, differential entropy (DE) features are extracted from the EEG signals using a short-time Fourier transform (STFT) with a 4 s nonoverlapping Hanning window. These features are divided into five frequency bands: δ (1-4 Hz), θ (4-8 Hz), α (8-14 Hz), β (14-31 Hz), and γ (31- 50 Hz). Thus, at every time step, we have DE features of 62 channels, each of which contains data in 5 frequency bands [1]. As a result, our EEG feature has the dimensions of 5×62.

For the EYE data, we adopt the same method used in [3] to extract the features. The features contain the information of pupil diameter, dispersion, fixation, saccade and blink. This information has been demonstrated to be useful in emotion recognition [3] and has been widely used in the literature [3], [4], [5], [6]. The EEG feature has dimensions of 41×1, and we repeat it to the size of 41×62 for concatenation with the EEG feature.

For the EIG data, they are recorded every 1 s and downsampled with a 4 s window. Then, EIG data are fed into the model directly. In our model, we use VGG19 [7] for feature extraction. We extract several middle layers outputs, reshape them into the same size and concatenate them to obtain V_0. For instance, we extract two middle layer outputs of the CNN with dimensions of $128\times20\times14$ and $256\times10\times7$. We reshape the former matrix into $512 \times 10 \times 7$ and concatenate them. Then, we obtain a V_0 with the size of $768 \times 10 \times 7$.

After features of all of the modalities are extracted, the EEG and EYE features are concatenated to obtain Q_0.

2) Stacked DCA Layers: After feature extraction, the concatenated feature of EEG and EYE, Q_0, is sized to $(d_q \times n_q)$, and the EIG feature, V_0, is sized to $(d_v \times n_v)$. Q_0 and V_0 are fed into the DCA layer directly.

Algorithm 1 Dense Co-Attention Layer

Input: the $(i-1)$-st DCA Layer output $Q_{i-1}(d_q \times n_q)$, $V_{i-1}(d_v \times n_v)$

Output: the i-st DCA Layer output $Q_i(d_q \times n_q)$, $V_i(d_v \times n_v)$

1: Compute $A_i(n_v \times n_q)$:

$$A_i = {V_{i-1}}^T W_i Q_{i-1} \tag{1}$$

where $W_i(d_v \times d_q)$ is a learnable weight matrix.

2: Compute $A_{Q_i}(n_v \times n_q)$ and $A_{V_i}(n_q \times n_v)$:

$$\begin{aligned} A_{Q_i} &= softmax(A_i) \\ A_{V_i} &= softmax({A_i}^T) \end{aligned} \tag{2}$$

3: Compute $P_{Q_i}(d_q \times n_v)$ and $P_{V_i}(d_v \times n_q)$:

$$\begin{aligned} P_{Q_i} &= Q_{i-1}{A_{Q_i}}^T \\ P_{V_i} &= V_{i-1}{A_{V_i}}^T \end{aligned} \tag{3}$$

4: Compute $Q_i(d_q \times n_q)$ and $V_i(d_v \times n_v)$:

$$\begin{aligned} Q_i &= ReLU(W_{Q_i}\begin{bmatrix} Q_{i-1} \\ P_{V_i} \end{bmatrix} + b_{Q_i}) + Q_{i-1} \\ V_i &= ReLU(W_{V_i}\begin{bmatrix} V_{i-1} \\ P_{Q_i} \end{bmatrix} + b_{V_i}) + V_{i-1} \end{aligned} \tag{4}$$

where $W_{Q_i}(d_q \times (d_q + d_v)$, $W_{V_i}(d_v \times (d_q + d_v))$ and $b_{Q_i}(d_q)$, $b_{V_i}(d_v)$ are learnable weight matrix and bias.

5: **Return** $Q_i(d_q \times n_q)$, $V_i(d_v \times n_v)$

The co-attention layer (DCA layer) is the key architecture that introduces the attention mechanism among the modalities. The DCA layer performs the computation of attended features, extracts different modality representations row/columnwise and transforms the representation into outputs with a single-layer MLP and residual connection [8].

The calculation process can be described in Algorithm 1, and the structure is shown in Fig. 2. For the DCA (i)-st layer, the output from the previous layer Q_{i-1} and V_{i-1} first

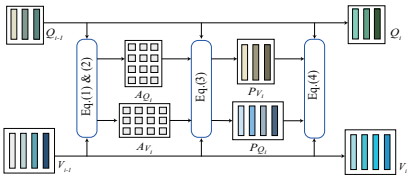


Fig. 2. Structure of the co-attention layer

interact by a multiplication with a learnable weight matrix W_i (Eq. (1)). Then, we use softmax along the row and column to obtain A_{Q_i} and A_{V_i} (Eq. (2)). Next, Q_{i-1} and V_{i-1} interact for the second time by multiplying A_{Q_i} and A_{V_i}, respectively, obtaining P_{Q_i} and P_{V_i} (Eq. (3)). Finally, the result of the DCA (i)-st layer, Q_i and V_i are computed after the last interaction by concatenating Q_{i-1} to P_{V_i}, V_{i-1} to P_{Q_i}, feeding into the MLP and a skip connection (Eq. (4)). In Eqs. (1), (3) and (4), modalities interact densely through multiplication with learnable matrices, skip connection and MLP.

Since the DCA layer does not change the dimensions of the input and output, we stack the DCA layers to form a hierarchy, strengthen the attention mechanism and build the correlation between the different modalities layer-by-layer.

For the dense interactions in one DCA layer and stacking of the DCA layers multiple times, the relationship and attention mechanism among modalities are gradually emphasized instead of the feature itself. Considering that the feature itself is not emphasized, we make an assumption that in the practical applications where it is difficult to collect the EEG signals, even randomly selected EEG data can also induce some of the relationship and attention mechanism with EYE and EIG data to eventually aid in recognition.

3) Prediction: After the calculation through the stacked DCA layers, all DCA layer outputs Q_i and V_i $(i = 0, 1, 2, \ldots, l)$ are fed into an MLP, followed by a softmax along the column to compute the attention weights α_{Q_i} and α_{V_i} $(i = 0, 1, 2, \ldots, l)$, respectively. The attention weights α_{Q_i} and α_{V_i} $(i = 0, 1, 2, \ldots, l)$ measure the contribution of the different dimensions and correlation among the modalities.

$$\begin{aligned} \alpha_{Q_i} &= Softmax(MLP(Q_i)) \\ \alpha_{V_i} &= Softmax(MLP(V_i)) \end{aligned} \tag{5}$$

We note that Q_i and α_{Q_i} have the same size $(d_q \times n_q)$, and V_i and α_{V_i} have the same size $(d_v \times n_v)$.

Then, we perform elementwise multiplication of Q_i and α_{Q_i}, V_i and α_{V_i} to obtain a weighted value, and we sum along the column, obtaining the vector $s_{Q_i}(d_q \times 1)$ and $s_{V_i}(d_v \times 1)$.

$$\begin{aligned} s_{Q_i} &= Sum(Q_i \times \alpha_{Q_i}) \\ s_{V_i} &= Sum(V_i \times \alpha_{V_i}) \end{aligned} \tag{6}$$

Finally, the outputs of every DCA layer Q_i and V_i $(i = 0, 1, 2, \ldots, l)$ are computed to obtain s_{Q_i} and s_{V_i} $(i = 0, 1, 2, \ldots, l)$. Then, we concatenate all s_{Q_i} and s_{V_i} $(i = 0, 1, 2, \ldots, l)$ and feed it into an MLP to predict the label.

TABLE I

BASELINE, ALL/TWO MODALITIES AND REEG ACCURACY

	EIG	EYE	EEG	ALL	TWO	REEG
Ave.	67.95	77.80	78.51	**87.63**	86.28	86.81
Std.	14.02	14.61	14.32	**6.45**	6.62	7.63

III. EXPERIMENTAL SETTINGS AND RESULTS

A. Dataset

We use the SEED[1] dataset, which contains three emotions: happy, sad, and neutral. Nine healthy subjects (4 males and 5 females) aged between 20 and 24 participate in the experiment [2]. Each subject watches 15 four-minute emotional film clips and performs the experiments three times, with an interval of approximately three or four days. Therefore, there are total 27 experiments in the dataset.

The EEG signals are recorded by an ESI NeuroScan System with 62 channels at a sampling rate of 1000 Hz. The EYE and EIG data are recorded simultaneously with SMI ETG eye tracking glasses [3], [4].

B. Experimental Settings

We use three modalities to recognize emotion first. For practical applications where it is difficult to collect EEG signals, we also evaluate our model only using EYE and EIG data.

In addition to recognition of emotion with three and two modalities, we propose a new classification paradigm that only needs to collect EYE and EIG data and uses randomly selected EEG signals with a random label that we already have to induce the attention mechanism. We name this new paradigm REEG because the EEG data with the corresponding label are replaced by random EEG data. This paradigm is proposed to improve the recognition performance with two modalities (EYE and EIG), where we can use EEG data that we have on hand instead of collecting EEG data. We denote our 27 experiments as $E_i(i = 1, 2, \ldots, 27)$. For experiment E_k, we only use the EYE and EIG data. For EEG data, we randomly select EEGs (random labels) from the other 26 experiments $E_i(i \neq k)$ and replace all the original EEG data of experiment E_k. We take the EYE and EIG data of experiment E_k together with randomly selected EEG data from experiment $E_i(i \neq k)$ as the three modalities to recognize emotion.

For performance comparison with the existing approaches, we use the same experimental setup with [3] for the three recognition modes mentioned above. Fifteen data clips are divided into 5 groups, and each group has three different emotions. Then, we take the first 3 groups as the training set and the last 2 groups as the test set. For each experiment, 5-fold cross-validation is adopted in performance evaluation.

C. Experimental Results

Table I lists the accuracy of the three baseline results and two modified DCAN results on the SEED dataset.

[1] http://bcmi.sjtu.edu.cn/~seed/index.html

In one-modality emotion recognition, the accuracy of the EEG and EYE features are from [3] with an SVM classifier. We input EIG data into a pretrained VGG19 [7] and then input the output into an SVM classifier, obtaining the accuracy of the EIG features. It is observed that the EEG features have the best performance, followed by EYE features, and the worst performance is obtained for EIG features. This means the EEG data have the best representation of human emotion compared to EYE and EIG data. Similar results have been reported in the literature [5].

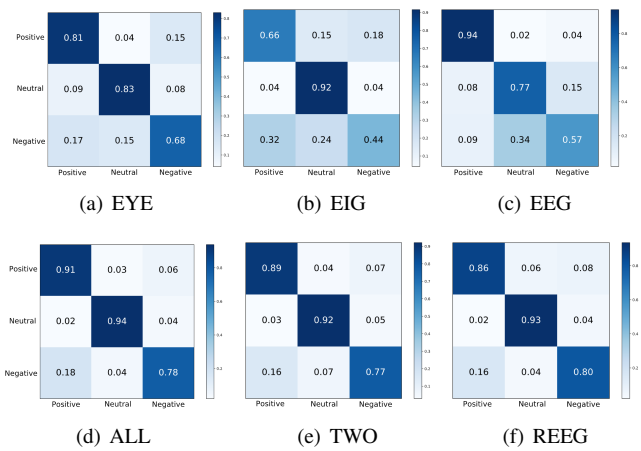
In three-modality emotion recognition (ALL), an accuracy of 87.63% and standard deviation of 6.45% are attained. Compared with three single-modality emotion recognition accuracy values in Table I, our performance is much more accurate and stable. This accuracy indicates that the attention mechanism is capable of characterizing the emotion state.

In two-modality (EYE and EIG) emotion recognition (TWO), an accuracy of 86.28% and standard deviation of 6.62% are achieved. The accuracy is lower than that of the three-modality recognition but is still much better than that of the single-modality approach. This result demonstrates that the use of information from only the eye can also achieve good accuracy and that this information can be used for emotion recognition when EEG data are unavailable.

The new paradigm (REEG) has an accuracy of 86.81% and standard deviation of 7.63%. This accuracy is lower than that of three-modality recognition but is superior that of the two-modality approach, indicating that randomly selected EEG data can still induce some of the attention mechanism to recognize emotions.

Fig. 3 shows the confusion matrices for emotion recognition. The confusion matrices show that the negative emotion state is difficult to recognize, while positive and neutral emotion states are relatively easier to recognize. Additionally, multimodal emotion recognition is better than recognition by one modality.

As observed in Figs. 3 (a), (b) and (c), EYE and EIG data predict the neutral emotion state most accurately, while EEG data predict the positive emotion state with 94% accuracy. This phenomenon confirms the results of previous studies that different modalities describe different emotion states [6]. From Figs. 3 (d)-(f), the accuracy of predicting neutral and negative emotion state is much higher than that of single-modality recognition, but the accuracy of predicting a positive emotion state is lower than that observed in Fig. 3 (c). This phenomenon shows that multimodal emotion recognition averages the effects of the three single modalities. Comparing Fig. 3 (e) and Fig. 3 (f), we find that when adding randomly selected EEG data, the accuracy of the predicted neutral and negative emotion state is improved. This fact indicates that randomly selected EEG data can help support the attention mechanism and thus recognition.



(a) EYE (b) EIG (c) EEG

(d) ALL (e) TWO (f) REEG

Fig. 3. Confusion matrices for emotion classification.

IV. CONCLUSION AND FUTURE WORK

In this paper, we have introduced a multimodal, attention mechanism-based emotion recognition model, a modified Dense Co-Attention Symmetric Network. With this model, we have investigated EIGs and the attention mechanism for multimodal emotion recognition. By using three modalities, the proposed model achieves a stable accuracy of 87.63%. With two modalities (EYE and EIG), our model obtains a stable accuracy of 86.28%, meaning that it is more suited for practical use. Moreover, we have proposed a new classification paradigm REEG that can improve the recognition performance of two modalities and does not require EEG data collection.

This work demonstrates the feasibility of recognizing human emotion with the attention mechanism. More importantly, the attention mechanism overcomes the bottleneck of requiring EEG signals for motion recognition, implying that less expensive and more convenient practical applications can be implemented.

ACKNOWLEDGMENT

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, the 111 Project, and the China Southern Power Grid (Grant No. GDKJXM20185761).

REFERENCES

[1] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu: Differential entropy feature for EEG-based emotion classification. In: IEEE NER, pp. 81–84 (2013).
[2] W.-L. Zheng, B.-L. Lu: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. TAMD, pp. 162–175 (2015).
[3] Y. Lu, W.-L. Zheng, B. Li, B.-L. Lu: Combining eye movements and EEG to enhance emotion recognition. In: IJCAI, pp. 1170–1176 (2015)
[4] W. Liu, W.-L. Zheng, and B.-L. Lu: Emotion recognition using multimodal deep learning. In: ICONIP, pp. 521-529 (2016).
[5] Y.-T. Lan, W. Liu and B.-L. Lu: Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In: IJCNN (2020).
[6] J.-J Guo, R. Zhou, L.-M. Zhao and B.-L. Lu: Multimodal Emotion Recognition from Eye Image, Eye Movement and EEG Using Deep Neural Networks. In: IEEE EMBC, pp. 3071-3074 (2019).
[7] K. Simonyan and A. Zisserman: Very deep convolutional networks for large-scale image recognition. In: ICLR, CoRR abs/1409.1556 (2015).
[8] D.-K. Nguyen and T. Okatani: Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: CVPR, pp. 6087-6096 (2016).