Multimodal depression detection using a deep feature fusion network

Guangyao Sun[a], Shenghui Zhao[a], Bochao Zou[b*], Yubo An[a]

[a] School of Information and Electronics, Beijing Institute of Technology, Beijing, China; [b] School of Computer and Comunication Engineering, University of Science and Technology Beijing, Beijing, China

ABSTRACT

Currently, more and more people are suffering from depression with the increase of social pressure, which has become one of the most severe health issues worldwide. Therefore, timely diagonosis of depression is very important. In this paper, a deep feature fusion network is proposed for multimodal depression detection. Firstly, an unsupervised autoencoder based on transformer is applied to derive the sentence-level embedding for the frame-level audiovisual features; then a deep feature fusion network based on a cross-modal transformer is proposed to fuse the text, audio and video features. The experimental results show that the proposed method achieves superior performance compared to state-of-the-art methods on the English database DAIC-WOZ.

Keywords: Depression detection, unsupervised learning, transformer, multimodal feature fusion

1. INTRODUCTION

Depression, also known as depressive disorder, is a major type of mood disorders. Currently, the main depression detection methods rely on specific questionnaires, such as the clinician-administered Hamilton Rating Scale for Depression (HAMD)[1]. These methods are time-consuming and labor-intensive. With the rise of deep learning, many studies use deep neural networks to process multimodal features for depression assessment. Yang et al.[2] fed the features of text, audio, and video into a DCNN network to obtain regression results respectively. Then outputs of the three networks were passed into a DNN network for decision fusion. Alhanai et al.[3] deployed a LSTM model to capture the temporal relationship between audio features and text features for depression detection. Lin et al.[4] used the LSTM and CNN to process text features and audio features respectively. Then the outputs of the two models were combined for late fusion.

In the field of automatic depression detection, current research has certain limitations. Firstly, the features extracted from audio and video are frame-level, and the existing methods to encode frame-level features into sentence-level use statistical functions which lead to the loss of the temporal relationship between frames. Secondly, most current network structures for multimodal fusion use decision fusion or late fusion which have limited ability to capture the complementarity of features between different modalities. In this work, we tried to overcome the aforementioned drawbacks. Our main contributions can be summarized as follow.

(1) An unsupervised autoencoder network is proposed to obtain sentence-level vector of frame-level audio and video features for depression detection.

(2) We propose a cross-modal transformer based deep fusion network for multimodal features to detect depression.

2. METHOD

2.1 Model overview

The overall structure of our proposed method is shown in Figure 1.

Firstly, the feature representation module is used to obtain sentence-level vectors of audio, text, and video. Then the feature fusion module based on cross-modal Transformer is adopted to combine three modalities. Next, a Bi-LSTM with self-attention is used to capture the temporal relationship of each single modality. Finally, the outputs of the self-attention module are fed into the low-rank fusion module for further late fusion to get the final regression result.

[*] zoubochao@ustb.edu

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066A
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2662620

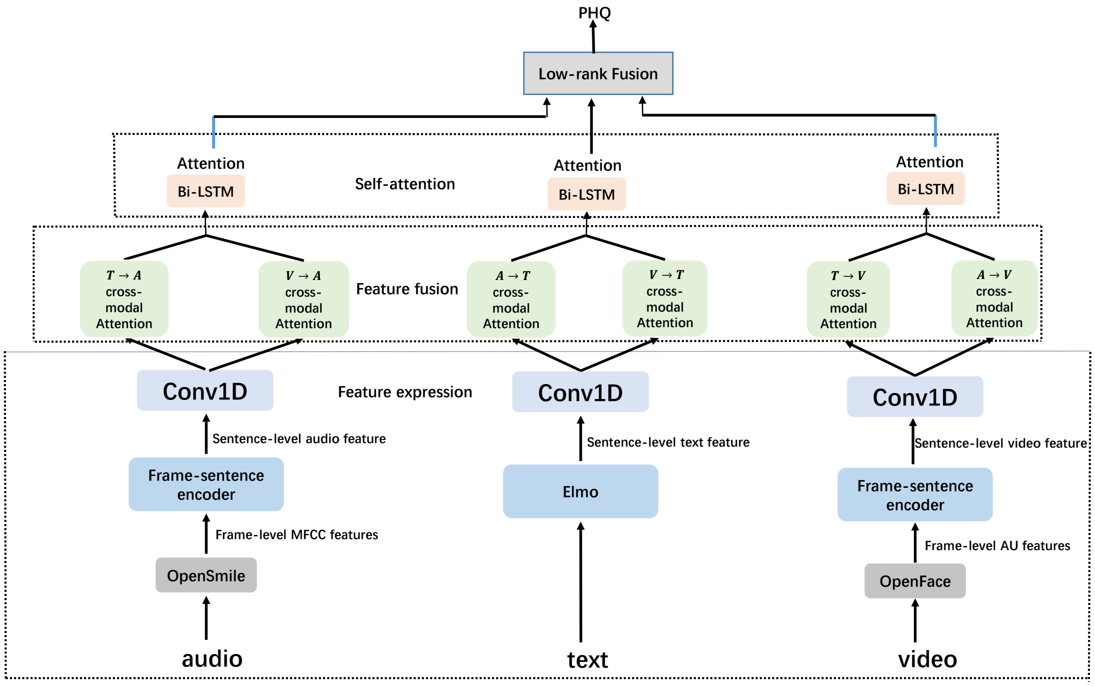
PHQ
Low-rank Fusion
Attention
Bi-LSTM
Self-attention
Attention
Bi-LSTM
Attention
Bi-LSTM
T → A cross-modal Attention
V → A cross-modal Attention
Feature fusion
A → T cross-modal Attention
V → T cross-modal Attention
T → V cross-modal Attention
A → V cross-modal Attention
Conv1D
Feature expression
Conv1D
Conv1D
Sentence-level audio feature
Sentence-level text feature
Sentence-level video feature
Frame-sentence encoder
Elmo
Frame-sentence encoder
Frame-level MFCC features
Frame-level AU features
OpenSmile
OpenFace
audio
text
video

Figure 1. An overview of the proposed model.

2.2 Feature extraction

The dataset consists of three modalities, namely text, video and audio features. For the text data, we use a pre-trained 1024-dimensional Elmo[5] sentence embedding to encode the transcribed subjects' responses for each question. For the audio data, 39-dimensional frame-level MFCC features are extracted for each audio segment. For the video data, 20-dimensional frame-level AU features are extracted for each video segment.

2.3 Unsupervised autoencoder

The extracted video and audio features are both frame-level. Temporal aggregations are needed to compress sentence-level vectors before feature fusion. Most previous methods encode the frame-level features into sentence-level with statistical functions which may lead to the loss of the temporal information between frames. To solve this problem, we propose an unsupervised autoencoder based on Transformer[6]. The overall structure of the autoencoder is shown in Figure 2.

The frame-level features are fed into the frame-sentence encoder to obtain the sentence-level vector of multi-frame features. The frame-sentence encoder consists of three layers transformer encoding units. The transformer encoding unit is illustrated in Figure 3.

The Positional Encoding module is used to add positional information to the frame-level features X. The specific calculation formula of positional embedding can be expressed by equations (1) and (2).

$$PE_{(pos,2i)} = sin\left(pos/10000^{2i/F}\right) \tag{1}$$

$$PE_{(pos,2i+1)} = cos\left(pos/10000^{2i/F}\right) \tag{2}$$

Here pos is the frame number, F is the dimension of the frame-level feature, I is the index of the features, and $2i$ represents the even index while $2i+1$ represents the odd index.

The frame-level features X are then added to the positional embedding PE to generate the input vector I. Then the Multi-Head Attention module is used to calculate the temporal relationship between frames. The relationship between vector A and input vector I can be expressed by

$$Q = I \times W_Q \tag{3}$$

$$K = I \times W_K \tag{4}$$

$$V = I \times W_V \tag{5}$$

$$A = Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where W_Q, W_K, W_V denote linear matrixes,Q, K, V denote the query vector, key vector, and value vector, respectively, and d_k is the normalization coefficient.

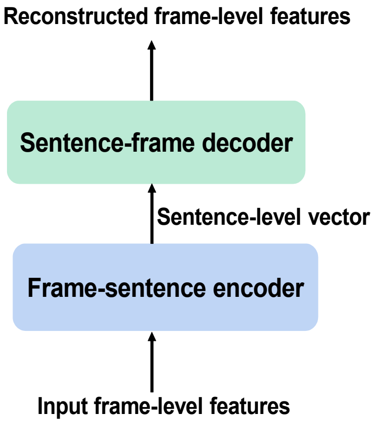

Figure 2. The structure of autoencoder.

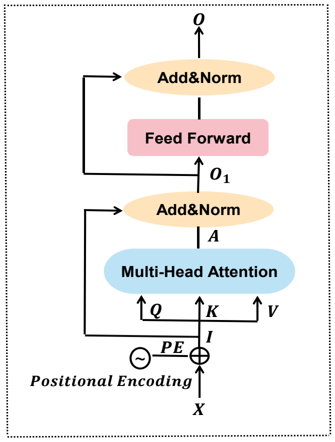

Figure 3. The structure of transformer.

Then vector A passes through residual connections and forward neural networks to obtain the output vector O. O is fused with the temporal and attention relationship between multi-frame features. The vector O of the last time step is the output of the frame-sentence encoder. After self-filling, a vector X' is passed through the sentence-frame decoder to reconstruct the frame-level features. The decoder is also composed of 3-layer Transformer coding units. The output vector Y of the decoder has the same dimension as the frame-level features X. The loss function of the unsupervised autoencoder is mean square error, and the difference between X and Y is calculated to update weights of the unsupervised autoencoder. When the model is converged, the output of the encoder is stored as a sentence-level vector representation of frame-level features.

So far, we have obtained three sentence-level vectors X_t, X_a, X_v of text, audio, and video modality respectively. Before the feature fusion module for deep fusion, three one-dimensional convolutional layers are utilized to compress the dimensions of the three sentence-level vectors. Then, $X_T \in R^{S\times30}, X_A \in R^{S\times30}, X_V \in R^{S\times30}$ are fed into the feature fusion module for deep fusion, where S is the question number of each subject and 30 is the dimension of compressed features.

2.4 Feature fusion

Six cross-modal transformers[7] are adopted to achieve a deep fusion of three modalities. We take $A \rightarrow T$ cross-modal transformer as an example to introduce the application of cross-modal transformer in the feature fusion module. The structure of $A \rightarrow T$ cross-modal transformer can be seen in Figure 4.

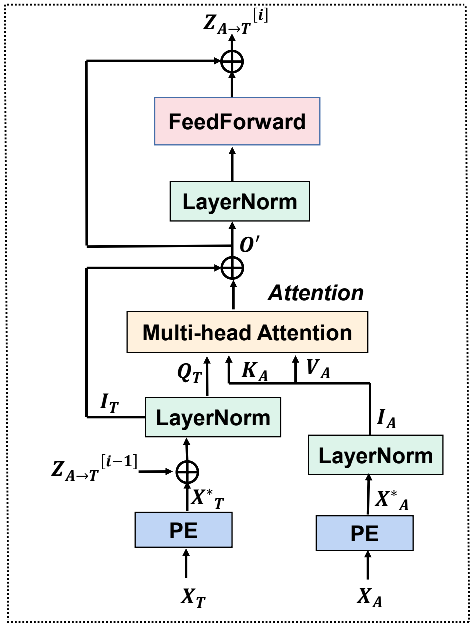

Figure 4. The structure of cross-modal transformer.

The relationship between the output of the i_{th} $A \to T$ cross-modal transformer $Z_{A\to T}{}^{[i]}$ and the input vector X_T, X_A can be represented by

$$X^*{}_T = PositionalEncoding(X_T) \tag{7}$$

$$X^*{}_A = PositionalEncoding(X_A) \tag{8}$$

$$I_T = LayerNorm\left(X^*{}_T + Z_{A\to T}{}^{[i-1]}\right) \tag{9}$$

$$I_A = LayerNorm(X^*{}_A) \tag{10}$$

$$Q_T = I_T W_{Q_T} \tag{11}$$

$$K_A = I_A W_{K_A} \tag{12}$$

$$V_A = I_A W_{V_A} \tag{13}$$

$$Attention = softmax\left(\frac{Q_T {K_A}^T}{\sqrt{d_k}}\right)V_A \tag{14}$$

$$O' = Attention + I_T \tag{15}$$

$$Z_{A\to T}{}^{[i]} = O' + FFN\left(LayerNorm(O')\right) \tag{16}$$

where Positional Encoding is used to add positional information to X_T and X_A. The specific relationship can be obtained from equations (1) and (2). LayerNorm is the layer normalization operation. $Z_{A\to T}{}^{[i-1]}$ is the output of $(i-1)_{th}$ layer cross-modal transformer. W_{Q_T}, W_{K_A}, W_{V_A}denote linear matrixes. And Q_T, K_A, V_A denote query vector of text modality, key vector, and value vector of audio modality. d_k is the normalization coefficient. FFN is the forward neural network which is composed of linear layers.

The outputs of the cross-modal transformer are concatenated to generate $Z_T \in R^{S\times 60}$, $Z_A \in R^{S\times 60}$, $Z_V \in R^{S\times 60}$. Z_T, Z_A, Z_V deeply fuse the information of the other two modalities and are then fed into the self-attention module to capture the temporal information of each modality.

2.5 Self-attention

Bi-LSTM with attention mechanism is utilized to capture the temporal relationship of Z_T, Z_A, and Z_V. Taking Z_T for example, the Bi-LSTM model with a hidden size of 30 is used to capture the temporal relationship between each time step of Z_T. Considering that features at different time steps are of different importance to the final assessment result, an attention mechanism is introduced to weight the output of Bi-LSTM at different time steps. The relationship between the output of self-attention module S_T and the output of Bi-LSTM $Z^{\circ}{}_T$ can be represented by

$$Z^*{}_T = tanh \times Z^{\circ}{}_T \tag{17}$$

$$alpha = softmax(w \times Z^*{}_T) \tag{18}$$

$$Z'{}_T = alpha \times Z^{\circ}{}_T \tag{19}$$

$$S_T = \sum_{i=1}^{T} Z'{}_T \tag{20}$$

where w is the weighting factor and T is the question number.

Then S_T is passed through a ReLU layer and a linear layer to obtain the output O_T of self-attention module. Finally, three temporal vectors O_T, O_V, O_A are passed through the late fusion module for further fusion to obtain the final assessment result.

2.6 Late fusion

The low-rank fusion[8] is used to further fuse O_T, O_A and O_V. The relationship between the final output H and O_T, O_V, and O_A can be represented by

$$\begin{aligned} H &= \left(\sum_{i=1}^{r} \otimes_{m=1}^{M} w_m{}^{(i)}\right) \cdot (z_l \otimes z_a \otimes z_v) \\ &= \bigwedge_{i=1}^{M}\left[\sum_{i=1}^{r} w_m{}^{(i)} \cdot z_m\right] \end{aligned} \tag{21}$$

where z_l, z_v, and z_a are the vectors O_T, O_V and O_A padding with 1, M is the number of modality 3, $\bigwedge_{i=1}^{M}$ denotes the element-wise product over a sequence of tensors, and W is the weight tensor. And the value of r is 17, which is the number of rank factors.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In the experiment, the evaluation metrics of the classification task are Precision, Recall, and F1. For the regression task, the evaluation metrics are MAE and RMSE. Our experiments are conducted on the DAIC-WOZ[9] database which has a training set (77 depressed, 30 non-depressed) for training the network, validation set (23 depressed, 12 non-depressed) to verify network performance during training and testing set (33 depressed, 14 non-depressed) for testing network performance. Three baseline models[2-4] are compared to our proposed network DDFN, which are introduced in the Introduction section. Experimental results are shown in Table 1.

The proposed deep feature fusion network achieves superior results in both regression and classification tasks. Compared to current state-of-the-art methods, the lowest RMSE of 5.35 and the highest Precision of 0.91 and F1 of 0.89 were obtained by the proposed deep feature fusion network DDFN. Overall, our proposed deep feature fusion network DDFN can deeply fuse the features of three modalities for depression assessment. Compared with decision fusion and simple feature fusion network, better evaluation results are obtained by our deep feature fusion network DDFN.

Table 1. Experimental results with DAIC-WOZ.

Methods	MAE	RMSE	PRECISION	RECALL	F1
DCNN+DNN[2]	5.16	5.97	/	/	/
LSTM[3]	5.13	6.50	0.71	0.83	0.77
LSTM+CNN[4]	3.75	5.44	0.79	0.92	0.85
DDFN (proposed)	3.78	5.35	0.91	0.88	0.89

4. CONCLUSIONS

In this paper, we propose an unsupervised autoencoder to encode frame-level features into sentence-level. A deep feature fusion network is proposed to further fuse features of three modalities for depression detection. The experiment results show that our network DDFN achieves improved performance compared to state-of-the-art methods. Considering the insufficient samples in the depression database, we will investigate multimodal few-shot learning for depression detection in the next step.

ACKNOWLEDGMENTS

This research is supported in part by National Natural Science Foundation of China (U19B2032) and Guangdong Basic and Applied Basic Research Foundation (2021A1515110249).

REFERENCES

[1] Hamilton, M. A. X., "Development of a rating scale for primary depressive illness," *Br. J. Soc. Clin. Psychol.* 6(4), 278-296(1967).

[2] Yang, L., Jiang, D., Xia, X., Pei, E.,, Oveneke, M. C. and Sahli, H., "Multimodal measurement of depression using deep learning models," *AVEC 2017,* 53-9(2017).

[3] Al Hanai, T., Ghassemi, M. M. and Glass, J. R., "Detecting depression with audio/text sequence modeling of interviews," *Interspeech 2018,* 1716-20(2018).

[4] Lin, L., Chen, X., Shen, Y. and Zhang, L., "Towards automatic depression detection: A BiLSTM/1d CNN-based model," *Applied Sciences,* 10(23), 8701(2020).

[5] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M. and Clark, C., "Deep contextualized word representations," *NAACL* 2018, 2227-37(2018).

[6] Vaswani, A., Shazeer, N., et al., "Attention is all you need," *arXiv:1706.03762, (*2017*).*

[7] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P. and Salakhutdinov, R., "Multimodal transformer for unaligned multimodal language sequences," *ACL 2019,* 6558-69(2019).

[8] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A. and Morency, L. P., "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064,* 2018.

[9] Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al., "The distress analysis interview corpus of human and computer interviews," *LREC 2014,* 3123-28(2014).