2024, 21(2 Special Issue): 0548-0560
https://doi.org/10.21123/bsj.2024.9454
P-ISSN: 2078-8665 - E-ISSN: 2411-7986

Baghdad Science Journal

Research on Emotion Classification Based on Multi-modal Fusion

Xiang zhihua*[1,2], ***Nor Haizan Mohamed Radzi***[1], ***Haslina Hashim***[1]

[1]Faculty of Computing, University Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia.
[2]Guangdong Technology college, 526100, Qifu Avenue, Gaoyao District, Zhaoqing City, Guangdong Province, China.
*Corresponding Author.

ICAC2023: The 4th International Conference on Applied Computing 2023.

Received 14/09/2023, Revised 10/02/2024, Accepted 12/02/2024, Published 25/02/2024

Abstract

Nowadays, people's expression on the Internet is no longer limited to text, especially with the rise of the short video boom, leading to the emergence of a large number of modal data such as text, pictures, audio, and video. Compared to single mode data ,the multi-modal data always contains massive information. The mining process of multi-modal information can help computers to better understand human emotional characteristics. However, because the multi-modal data show obvious dynamic time series features, it is necessary to solve the dynamic correlation problem within a single mode and between different modes in the same application scene during the fusion process. To solve this problem, in this paper, a feature extraction framework of the three-dimensional dynamic expansion is established based on the common multi-modal data, for example video , sound ,text.Based on the framework, a multi-modal fusion-matched framework based on spatial and temporal feature enhancement, respectively to solve the dynamic correlation within and between modes, and then model the short and long term dynamic correlation information between different modes based on the proposed framework. Multiple group experiments performed on MOSI datasets show that the emotion recognition model constructed based on the framework proposed here in this paper can better utilize the more complex complementary information between different modal data. Compared with other multi-modal data fusion models, the spatial-temporal attention-based multimodal data fusion framework proposed in this paper significantly improves the emotion recognition rate and accuracy when applied to multi-modal emotion analysis, so it is more feasible and effective.

Keywords: Dynamic correlation, Feature matching, Multi-modal emotion classification, Match fusion, Temporal attention.

Introduction

Since the 21st century, information network technology has developed rapidly, and the upsurge of short videos with the participation of the whole people makes Internet information present a trend of diversification. This trend has attracted the attention of multi-modal emotion research and also triggered a wave of research upsurge. In recent years, the analysis and research results of the implied emotional tendency of multi-modal data have shown good value in the application process of Product marketing and recommendation, Health monitoring, and other aspects.

Earlier studies mainly used single-modal data for emotion classification, such as pictures and text modes[1].In recent years, the academic research findings on sentiment classification mainly focus on deep learning-based multi-modal technology to improve the results in emotion classification[2]. However, the external expression and perception of emotions have individual differences and various forms, influenced by the outside world. It is sometimes subjectively hidden or disguised, so the features are difficult to capture. In some studies, physiological signals, facial expressions, gestures, and other features are comprehensively used to study emotion classification. The physiological responses of different emotions are also different. Most traditional emotion recognition research focuses on feature analysis and feature fusion between a single mode or two modes, and emotion recognition is carried out in combination with expressions, gestures, physiological signals, and other ways. In recent years, researchers are no longer limited to the information analysis and fusion of single or two modes, but to further study the correlation between multi-modal information.

Related Studies

1. Emotion classification

The traditional algorithm mainly focuses on text and images[3]. The task of text emotion classification is to mine the emotion polarity contained in each text. Text emotion classification can be divided into emotion dictionary-based classification and machine learning-based emotion classification. At present, the use of deep learning technology for emotion analysis has also become the consensus of the academic community.

For the emotion classification process, First, build the emotion dictionary based on the prior expert knowledge, and then construct the mapping of emotion words with the corresponding emotions. After that, the text-matching method is used to identify the emotional statement contained in the text. Finally, the emotional tendency of the text is judged according to the emotional polarity of the emotional words in the text.

The main algorithms of emotion classification using traditional statistical machine learning at present are naive Bayes, maximum entropy algorithm, and so on. The main advantages of traditional machine learning methods include strong fitting ability, strong algorithm generalization ability, and overcoming the problem that the traditional emotion classification method based on an emotion dictionary cannot make good use of external vocabulary. However, it is difficult for traditional machine learning algorithms to extract good enough characteristics from raw data, thus limiting the performance of the algorithms.

Emotion analysis based on image data is also a hot topic of research in recent years. Its research is mainly based on facial expression data, which can be either static images or videos. Compared with static images, dynamic videos contain temporal data, which often contain more information. In the process of using this method for emotion analysis, it is necessary to locate the face position, cut the face image, and standardize the face image according to different factors such as light Angle, light and darkness, and horizontal deflection. At present, the analysis and extraction of face modal data in static images are mainly based on geometric features and texture features. As for the modal data of human face expression in the motion video, the common feature extraction methods include the optical flow method, the local binary mode of three orthogonal planes, etc.

Compared with image and video, there is noise in the speech signal, and the unity of the emotion labeling is insufficient. These problems make the processing of speech much more complicated. Direct emotional classification for long sound segments often leads to scattered and chaotic output results, with poor effect. Therefore, it is generally believed that the speech segment is needed in the process of emotion classification. At present, the segmentation of speech segments is usually divided into two methods: non-overlapping and overlapping segmentation. After segmentation, the existing speech emotion recognition method is used to extract emotional features and form a group of highdimensional speech emotional feature sets, which can obtain good performance.

2. Multi-modal Research

With the deepening of research, multi-modal emotion recognition has entered the vision of scholars and gradually expanded. Some representative research results have emerged in recent years.

The multi-modal fusion method was first divided into two categories: early fusion and post-fusion[4]. Early fusion is also known as feature-based fusion. Its principle is to integrate the features of each mode after extracting the data of each mode, and finally fuse the extracted multi-modal features to realize emotion analysis. The post-fusion is to make emotional predictions based on each mode of data, and then by voting, weighting, etc., make the final emotional judgments based on the outcome of each prediction. Because late fusion uses an independent model to model each modal data, therefore in the application process, it gives the application personnel a greater degree of freedom, but late fusion ignores the interaction between different modal data, therefore, it is difficult to deal with the complex multi-modal data analysis problems.

The current deep neural network modeling methods are used for multi-modal data, a bidirectional conversion model between modes, and a multi-modal data fusion model based on tensor. Some researchers have extended the traditional LSTM to uni-modal data, dividing the internal storage units and gates of the LSTM into different parts to characterize the multi-modal data. Some researchers used LSTM to extract context single-modal features to model the dependency of features at different moments in different modes, and then each uni-modal feature was taken as a multi-modal feature and entered into another LSTM network for the final emotion classification task. The above methods establish cross-modal interaction relations from multi-modal sequence data. However, similar to the shortcomings of traditional cyclic neural networks, these multi-modal emotion classification models based on the extension of cyclic neural networks will also suffer from performance degradation as the length of input sequence data increases. Zadeh et al. proposed the multi-attention cyclic network, which finds the dynamic dependency between different modes through multiple attention modules, and stores the matrix representation of the dependency in the hybrid storage unit of LSTM[5].

Compared with uni-modal and bimodal approaches, multi-modal emotion fusion requires considering more complex content. Since the single-mode method does not need to consider the advantages and disadvantages of fusion, the multi-mode emotion recognition, whether feature layer fusion or decision layer fusion, has to fully consider the computing relationship between multiple groups of vectors, which greatly increases the complexity, and has high requirements on the flexibility and reliability of the fusion algorithm.

Multi-modal Fusion-matched Network Based on Three-Dimensional Dynamic Expansion

1. Spatial and Temporal Feature-Enhanced Network Based on the Three-Dimensional Dynamic Expansion

Multi-modal data shows obvious dynamic time series features. On one hand, multi-modal research requires complete feature extraction, and reasonable calculation modeling of the dynamic association in the time series data of a single mode, that is, the dynamic association within the mode. On the other hand, the outer network must grasp the global changes of the whole data, and compute and model the dynamic association information between different mode time series data, that is the dynamic association between modes.

In this paper, three mainstream multi-modal data of text, audio, and video are taken as the research object. First, extract the corresponding features from the multi-modal data, and then perform a vectorized representation of the text through the word embedding model, with the words in the text as the basic unit of alignment, obtain the vector representation about the specific modal data via the unsupervised learning algorithm. When dividing, it is based on the duration of the sound effect of each word. According to the division situation, the data of audio and video are also divided and extracted according to the same interval, to realize the alignment of the three modal data in the time series, and the corresponding features are also extracted from them. For the audio data, a basic two-dimensional network was used to extract the acoustic features, and for the video data, the underlying Three-Dimensional(3D) network was used to extract the visual features.

The processing based on visual information is the focus of the research, and the key is to explore the correlation between visual and other information. Due to the complete temporal dynamic-dependent information that exists in the multi-modal data representation, including the intra-modal and inter-modal dynamic correlations, there is an inherent time-series dynamic alignment between different modes. Over time, the dynamic correlations between the different modes provide more complex complementary information over the temporal and mode domains. To better utilize the multilevel features of images, it is necessary to use both Two-Dimensional(2D) and 3D information and establish the temporal correlation between them. Considering this multi-modal complementary information, the long-term feature changes of the poly-tropic data can be analyzed more accurately.

Visual attention is a creative method of visual algorithms and a basic network of multi-modal space-time attention. In the structure of CNN, the output of each layer is composed of several feature maps, and each "pixel" of these feature maps represents the response of the local region corresponding to the upper input to a specific convolution kernel. However, the contribution of features at each position to the final result is not the same, except that the gradient back-propagation guides the updating of model parameters to adjust the weight of each place. Therefore, the core parameters of each layer, namely the forward calculation, must be considered in the first stage, and the attention weight of each part must be annotated on the input amount of that stage.

In the specific processing, for 2D image data, corner points are usually used to represent some gradient mutation points, which are the areas that need to pay the most attention. The Harris corner point moves in any direction (u, v) and will change significantly. Harris corner detection is a common corner detection algorithm. It screens some points with drastic changes through the first step of the image. Human eyes are more sensitive to corner points with drastic changes in the surrounding grayscale, and drastic changes mean that the response function value of this local position is large, whereas the response value is small[6]. This article proposes a dynamic extended feature enhancement modular Harris corner algorithm. It filters out points with drastic changes through the first step of the image.

For the processing of video sequences, a 3D dynamic expansion algorithm is proposed to achieve specific results. It is shown in Fig.1.

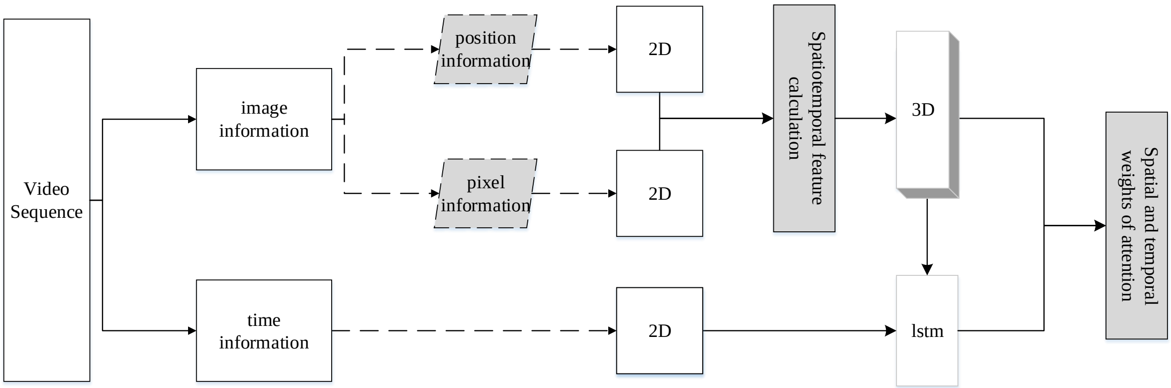

Figure 1. Basic Network for 3D Dynamic Expansion.

In this case, the pixel gradient can be calculated not only through the x and y axes but also along the t-axis of the time axis, that is, the spatial information can be extended to the time series to obtain the corner points on the 3D data. The Windows used by

Harris Corner to calculate the horizontal and vertical gradients are:

$$Tx = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \qquad 1$$

$$Ty = \begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \qquad 2$$

The gradient of the image in the x and y directions can also be calculated based on the horizontal and vertical windows:

$$I_x = I \otimes Tx, I_y = I \otimes Ty \qquad 3$$

The gradient matrix M and the corner response value R of the image are obtained：

$$M = \begin{bmatrix} I_x^2 & I_x \times I_y \\ I_x \times I_y & I_y^2 \end{bmatrix} \qquad 4$$

$$R = \det(M) - k \times (trace(M)^2) \qquad 5$$

Where det(M) is the determinant of M, trace(M) is the locus of M, and K is a constant. The response value of corner points reflects the possibility that each pixel is a corner point. By local non-maximum suppression, the maximum value in the local area is selected to obtain alternative corner points, and other values are removed. Then, the maximum value is compared with a threshold value, and the final corner position is screened, which is the most obvious gradient change in the image. Finally, several key point locations in 3D data Spaces are obtained by extending the 2D key point detection algorithm to 3D.

$$M_3 = \begin{bmatrix} I_x^2 & I_x \times I_y & I_x \times I_t \\ I_x \times I_y & I_y^2 & I_x \times I_y \\ I_x \times I_t & I_y \times I_t & I_t^2 \end{bmatrix} \qquad 6$$

$$R_3 = \det(M_3) - k(trace(M_3)^3) \qquad 7$$

Where, I_t is the gradient of input data on the time axis, and M_3 is the corresponding 3D focus value. The transformation of data in vector space is the feature representation of different directions of data. The responses of these positions were suppressed by local non-maximum value, and the key points were obtained after the small responses were eliminated. Through such judgment and screening, the important position data expression was obtained. To better show the importance of this information, a 3D vector is constructed with the core location data as the center. After the location information of these key areas is obtained, the response values of these locations are added to the original all-in-one weight graph to enhance the attention of this area. Finally, an attention-weight graph consistent with the original input data is obtained.

The number of channels of the feature weight map calculated from the general network is the number of frames of the input image sequence, but the number of channels output per layer is much larger. If this problem is solved, it facilitates the subsequent calculation. After merging the weight diagram, the convolution layers can be introduced to perform the convolution operation. In this process, the function of the convolution operation is to initialize its parameters and fuse them, and before this process, the number of channels in the feature weight map is preferred to be adjusted, so that the output of the corresponding convolution layer module remains consistent. Moreover, considering the complexity of the multi-modal data, it can be further optimized and tuned by introducing the attention layer, which helps to maintain the flexible implementation of the multi-modal data fusion framework based on attention matching in different scenarios.

2. Muti-Modal Fusion-Matched Network Based on Spatial and Temporal Feature Enhancement

After enhancing the spatial-temporal features based on 3D dynamic expansion, it is necessary to describe the 3D features deeply, To further calculate and analyze a large number of obtained feature groups. As time goes by, the dynamic correlation information between different modes provides more complex complementary information over the time and mode domain, So it is necessary to conduct matching modeling of the cross-modal and cross-temporal dynamic relationship between modes.

First, the feature matrix containing single-mode dynamic correlation information is connected in pairs, and all possible short and long-term interactions between the two modes are modeled through the coupling matrix in the double-linear transformation to obtain the corresponding matching fusion representation. Among them, the coupling matrix in the bilinear transformation can solve the inconsistent shape of the input matrix, thus ensuring that the coupling between different modes is more flexible.

$$M_{xy} = E_x W_1 (E_y)^T \in R^{t_x t_y} \quad 8$$

$$M_{yz} = E_z W_2 (E_y)^T \in R^{t_y t_z} \quad 9$$

Where, E$_x$, E$_y$, E$_z$ stands for single mode feature matrix; W$_1$, W$_2$ is the coupling matrix in a bilinear transformation; M$_{xz}$, M$_{yz}$ is the basic matching fusion matrix of the current modes xy and yz. The dimension of the fusion matrix corresponds to the current pattern and the time series, i.e. the fusion matrix is the mapping and match of the pattern features between the temporal directions. In this way, the feature vectors of modes at different moments do not jump and change with time, and the correlations between modes can be established through the time series.

The pre-temporal feature enhancement helps determine the feature importance of the current modes at the current time, that is, the modal matching fusion matrix optimizes the joint matrix representation of the two modes, by calculating the relative importance of the feature vectors at each moment. The length of the attention vector calculated by the attention mechanism is consistent with the length of the specific modal time series. Each element in the attention vector can measure the importance of the features at the corresponding moment. By assigning the attention weight to the matching fusion matrix, the model can focus on the more important fusion representation, and the calculation is shown in Formulas (10) and (11).

$$R_x = \tanh(W_x \cdot M_{xy}^T + b_x), R_z = \tanh(W_z \cdot M_{yz}^T + b_z)$$
10

$$\alpha_x = softmax(w_x \cdot R_x), \alpha_z = softmax(w_z \cdot R_z)$$
11

In the process of fusion, the attention weights need to be calculated. Usually, the matrix elements of the resulting tensor according to the mapping relationship, are fused and assigned, to obtain the attention fusion feature vector with a deep spatial-temporal relationship. The specific process is shown in Formulas (12).

$$M'_{xy} = broadcast(\alpha_x, M_{xy}), M'_{yz} = broadcast(\alpha_z, M_{yz}) \quad 12$$

Since the length of the attention vector is α$_x$, and the length of the first dimension of the matching fusion matrix is *M$_{xy}$*, follow $M'_{xy}[i,j]=\alpha_x[i]\cdot M_{xy}[i,j]$, where each element in the attention vector can be allocated directly in the form of broadcasting, So using this form of broadcast allocation attention weight, the size of the original matching fusion matrix can be maintained, that is, the size of *M$_{xy}$*, *M'$_{xy}$* is the same.

The multi-mode fusion-matched process can be regarded as the result of matrix operation based on crossover mode. The softmax function used to calculate the attention vector can normalize the output vector, so that all of the values of the elements in the vector are converted into numbers in the interval (0,1), and the values of all the elements add up to 1. This makes some elements of the attention vector close to zero. When assigning attention vectors to the modal matching fusion matrix, some values of the elements in the output are close to 0, making the matching fusion matrix highly sparse. For example, the length of the second dimension of the matching fusion matrix is consistent with t_y, the time series length of mode *Y*, indicating that the matching fusion matrix has the time information of mode *Y*. By matching the expansion operator of the fusion matrix, and then using the common time information of the mode Y in the matching fusion matrix, the matching from dual-mode to multi-mode can be realized, and the tensor P is shown in Formulas(13).

$$P = M'_{xy} \oplus M'_{yz} \in R^t \quad 13$$

This approach has two advantages. The core of the tensor operator is addition, which helps to avoid the higher sparsity of the tensor results obtained from the fusion of two matching matrices. On the other hand, the multi-modal tensor *P* has three dimensions, the first dimension length t_x, corresponds to the time series length t_x of mode *X*, the second pattern length is t_y, corresponds to the time series length t_y of mode *Y*, the third pattern length t_z, corresponds to the time series length t_z of mode *Z*. In multi-modal data fusion, each module eigenvector can match all feature vectors of the other two modules, conducive to modeling short-time dependence relationships. Finally, the attention vector is calculated again, and each element in the attention vector is allocated to obtain a tensor p'.

$$R_z = \tanh(W_z \cdot p + b_z) \tag{14}$$

$$\alpha_z = soft\max(w_z \cdot R_z) \tag{15}$$

$$p' = broadcast(\alpha_y, p) \tag{16}$$

Where $P'[i,j,k] = \alpha_y[k] \cdot p[i,j,k]$.

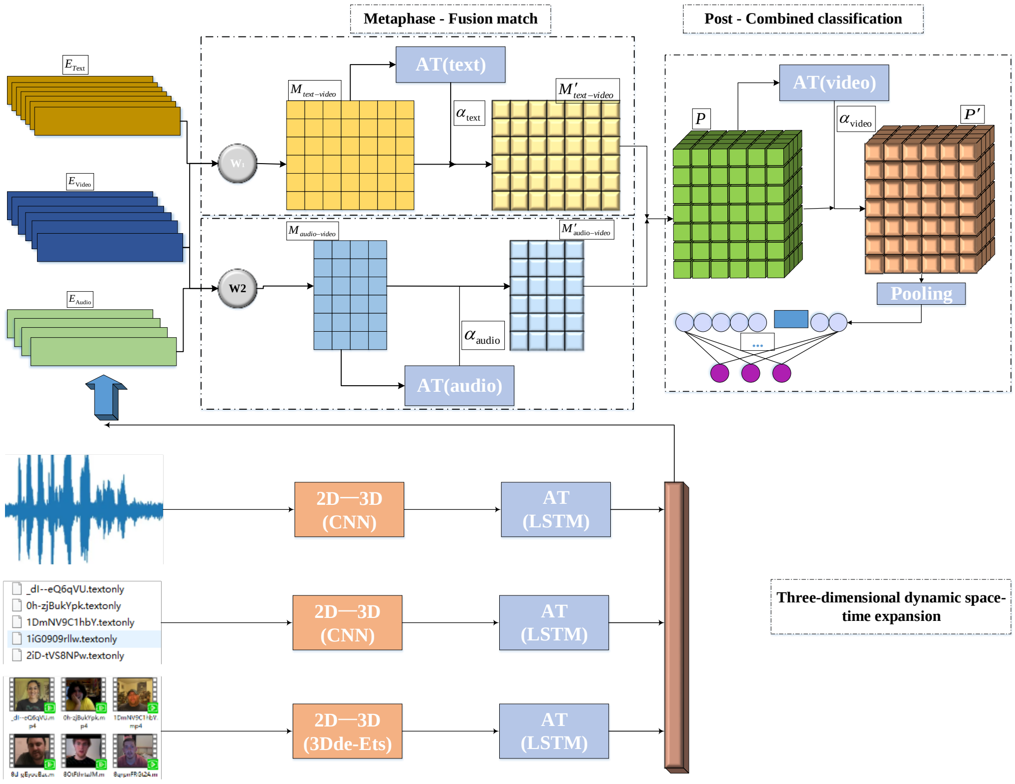

Figure 2. Multi-modal fusion-matched model based on spatial-temporal Feature Enhancement.

Therefore, the enhanced 3D features proposed earlier in the paper can be well matched with the fusion-matched model, and the multi-dimensional time accumulation avoids bringing higher sparsity to the tensor results obtained by the fusion-matched matrix. In the fusion of multi-modal data, the feature vectors of each mode can be matched with all the feature vectors of other modes. The framework is shown in Fig.2.

Experiments and Preparation

Datasets

MOSI: It is the mainstream data set for multi-modal emotion classification such as expression, speech, and text. It contains videos of 89 people(41 females and 48 males), who are mainly in their 20s to 30s. This data set first divides each video into multiple short videos based on one sentence, and uses these short videos as samples, from which to generate speech and text samples from the short videos. The labels are mainly marked in the form of scores, usually integers between -3 and 3. Among them, -3 means very negative, -2 means relatively negative, -1 means negative, 0 means neutral, 1 means positive, 2 means relatively positive, and 3 means very positive. Based on this criterion, the sample data were divided into negative samples (emotional results [-3,0)) and positive samples (emotional results (0,3]).

Experimental Setup and Evaluation Index

In the training stage of emotion recognition, the L1 loss function and Adam optimizer were used to optimize the trained emotion recognition model. The initial learning rate was set to 0.15, the number of iterations to 2000 epochs, and the batch size to 64. When the loss value of the loss function L1 loss is reduced to the minimum, the emotion recognition model at this time is saved, and the optimal model is used in the test to evaluate the performance.

Table 1. Experimental environment parameters

Experimental Environment	Configuration
CPU	I9-10850K
RAM	512G
GPU	Tesla T4
MB	16G
programming language	Python3
Word Embedding	BERT

The performance of multi-modal emotion recognition was comprehensively evaluated. The classification of emotion recognition is measured by 2-classification weighted accuracy, 2-classification F1 score, 3-classification weighted accuracy, and 7-classification weighted accuracy, while the performance of emotion recognition is measured by regression of the average absolute error MAE. Weighted accuracy is defined as follows:

$$WAccuracy = \frac{1}{n}(\sum_{i=1}^{n} s_i(y_i, \hat{y}_i)) \quad 17$$

$$s_i(y_i, \hat{y}_i) = \begin{cases} 1, y_i = \hat{y}_i \\ 0, y_i \neq \hat{y}_i \end{cases} \quad 18$$

Where, y_i is the true value, and $\hat{y}_i$ is the predicted value. The greater the weighted accuracy, the better the effect of emotion recognition.

Performance indicators include : Precision(P)-accuracy rate , Recall(R)-recall rate. F1-score(F1)-balanced F Score

$$p = \frac{TP}{TP + FP} \quad 19$$

$$R = \frac{TP}{TP + FN} \quad 20$$

$$F1 = 2 \times (\frac{\Pr ecision \times \mathrm{Re}\,call}{\Pr ecision + \mathrm{Re}\,call}) \quad 21$$

True positive (TP): Observation is predicted positive and is actually positive.

False positive (FP): Observation is predicted positive and is actually negative.

False Negative (FN): Observation is predicted negative and is actually positive.

TN (True negative): Observation is predicted negative and is actually negative.

MAE——The average Absolute Error between the actual value and the predicted value:

$$MAE = \frac{1}{m}(\sum_{i=1}^{m} |(y_i - \hat{y}_i)|) \quad 22$$

Results and Discussion

Ablation experiments

(1)2-classification
To verify the influence of multi-modal fusion on the accuracy of emotion classification, the modal features of different combinations of single mode (T, A, V), double mode (V+T, T+A, V+A), and three modes (T+V+A)were input to conduct the emotion classification experiment, and the

experimental results were compared respectively. The accuracy comparison of 2-classification is shown in Fig.3, and the F1 comparison is shown in Fig.4.

For single-mode information features, video information is processed by enhancing spatiotemporal features, and then directly used for analysis of emotional tendencies. For bimodal analysis, the information from different modes is first fused in pairs and then processed in the same way for emotional tendency analysis. For the third mock examination information features, the obtained two model features T+A, T+V, and V+T are fused to obtain the third mock examination fusion T+A+V, and the same processing method is used to classify emotions.

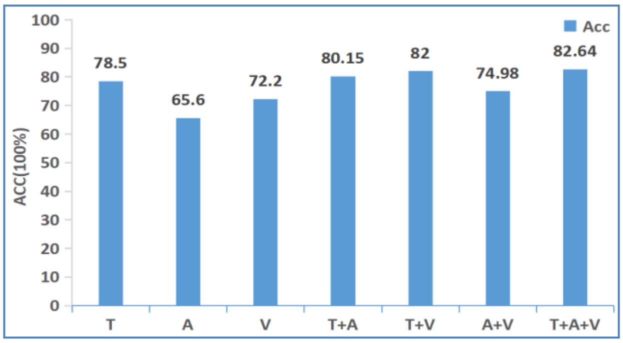

Figure 3. Accuracy comparison of 2-classification

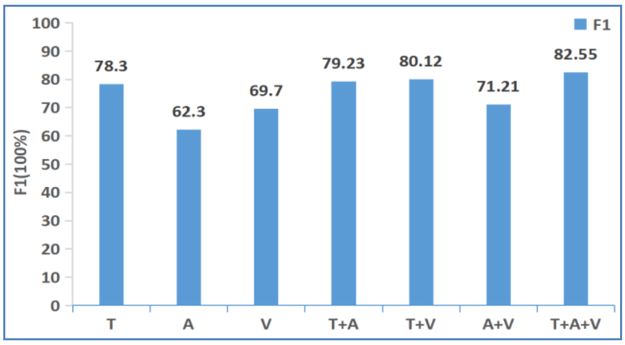

Figure 4. F1 comparison of 2-classification

By analyzing the data in Fig.3 and Fig.4, we can draw the following conclusions.

1) Among the temporal and spatial feature emotion classification using the attention mechanism, the accuracy of text is the highest, reaching 78.5%, and its F1 value is also the highest(78.3%). Speech, as a pattern type that is difficult to recognize in multi-modal emotion classification, still has lower accuracy and F1 value than video and text. It can be seen that after strengthening the three-dimensional expression of time and space, the feature expression of video signals will be stronger, and the emotion recognition rate of video signals reaches 72.2%, which has significantly improved.

2) Multi-modal emotion classification for video and text performed best with an accuracy of 82% and an F1 value of 80.12%, while the effect of audio and video is relatively low. The main reason is that the text is relatively stable, which can pull other modes, so text and video are easier to achieve unity in the timeline, It is easier to coordinate and assist the judgment of the emotional state of the current time node, realize the current hierarchical fusion of time and space. It can be found that the effect of each group of 2-modal emotion classification is better than the corresponding single-modal. Therefore, introducing another in one mode can indeed achieve the effect of feature complementation in the time series.

3) By integrating the features of facial expression, speech, and text, the accuracy rate and F1 value of the 3-modal emotion classification reached 82.64% and 82.55% respectively, which was improved compared with the 2-modal. The validity and feasibility of 3-modal emotion classification are validated, and the feasibility and effectiveness of the multi-modal emotion classification method proposed in this paper are also demonstrated.

(2)3-classification

The accuracy comparison of 3-classification is shown in Fig.5, and the F1 comparison is shown in Fig.6. As shown in Fig.5, the overall accuracy of the 3-classification is significantly lower than that of the 2-classification, because with the increase of emotion categories, the special features of "similar" affective expression overlap, and some features are almost the same. From the experimental results, the text and expression still get good results.

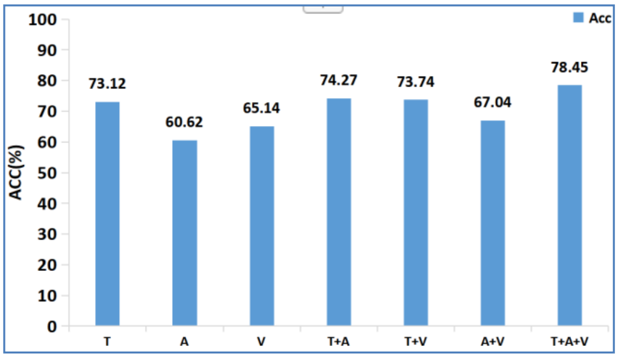

Figure 5. Accuracy comparison of 3-classification

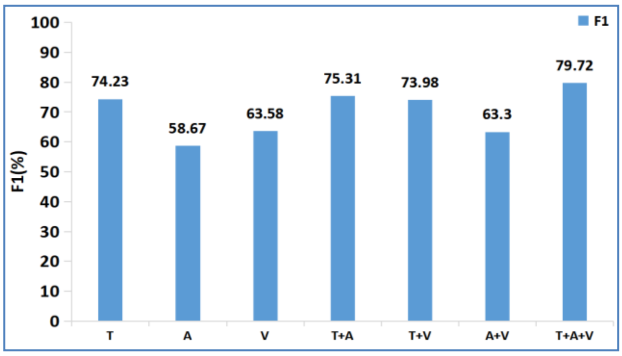

Figure 6. F1 comparison of 3-Classification

(3)7-classification

As shown in Fig.7, the accuracy rate of 7-classification reaches 37.76%, which proves that the temporality of spatial features of all modes favors stability under long-term changes.

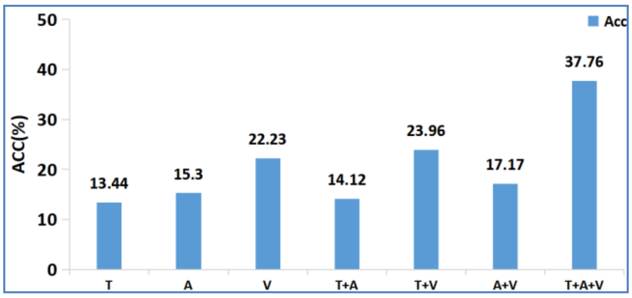

Figure 7. Accuracy comparison of 7-classification

Contrast experiment

In response to issues such as incomplete information and strong interference in single-mode emotion recognition, the latest four classic deep learning algorithm was selected for level comparison experiments. They are Memory Fusion Network (MFN), Recurrent Attended Variation Embedding Network(RAVEN), Multi-modal Cyclic Translation Network Model(MCTN), and Tensor Fusion Network(TFN).

MFN: This algorithm aligns multiple modalities (text, images, audio) very closely in the temporal dimension because data is extracted from different view information in the same video[7]. Due to the natural temporal modeling characteristics of LSTM, the MFN algorithm is introduced relatively smoothly. But for other cross-modal problems, such as Image Capture, Cross-modal retrieval, and Visual question answers, although different modalities are interconnected, there is no one-to-one alignment of fragments in the temporal dimension, and the application analysis of the cross-modal problem is not effective.

RAVEN: It is divided into three modules. For the word 'sick', there is a corresponding continuous video and audio, and the existing feature extraction tools are used to extract features[8]. Then the corresponding feature representations for each modality are obtained. Red represents video features, yellow for audio features, and green for word features. Calculate a score using video, audio, and word representations, and fuse the features based on this score to obtain a non-linguistic offset vector (purple). Finally, the vector is normalized and added to the word vector to obtain a word representation incorporating multi-modal information.

MCTN: It aims to learn a robust joint representation by converting between different modes, it can only use text modal data and create new latest results during testing. MCTN is end-to-end training with coupled translation prediction losses(including cyclic translation loss and predicted loss), to ensure that the joint representation of learning is task-specific (i.e. Multi-modal sentiment analysis)[9]. The advantage of MCTN is that once multi-modal data is used for training, only data from the source modality at the time of testing are required to infer the joint representation and labels. Therefore, MCTN is completely robust to test time disturbances or missing information in other modes.

TFN: In the encoding phase, the TFN encodes the input of the text modality using a network with LSTM and two fully connected layers, and encodes the input of voice and video modality using a 3-layer DNN network[10]. In the mode fusion phase, the output vector operates an external product operation after the encoding of the 3 modes, to obtain the multi-modal representation vector containing uni-modal information, double mode, and triple mode fusion information, for the next decision operation.

By comparing and analyzing the above models with the model proposed in this paper, it can be seen that the MFN model effectively solves the temporal synchronization of different modalities. However, no solution was considered for different modal data structures and cross-modal issues. Therefore, in the process of introducing video, audio, and other multi-modal data for comprehensive analysis, accuracy is bound to be affected; In the specific operation of the RAVEN model, the consideration of the internal features of the different modal is ignored, which leads to the phenomenon of over-fitting; The MCTN and TFN models both consider the temporal features of multi-modal data, and the reason why TFN has good robustness against test time disturbances and missing information in other modes depends on its human intervention during the testing phase. The algorithm model proposed in the article effectively solves the shortcomings of the above models to a certain extent, and theoretically, can improve the accuracy of multi-modal emotion classification.

The comparison of the above models is shown in Fig. 8. The MCTN is based on the structure of the encoder and decoder, learning the conversion relationship between modalities, exploring the connections between modalities, and achieving a classification accuracy of 79.3%. Compared to MCTN, the network in this paper has increased by 3.34%. The MFN multi-view gated storage network obtained multi-modal fusion features through time learning, with an accuracy rate of 36.2%. This further demonstrates the importance of temporal features in multiple classifications. The proposed method achieves a 1.56% improvement in accuracy compared to mature models such as MFN, because the proposed method not only explores the temporal correlation of modes but also reduces the problem of unmatched representation distributions of different modal features.

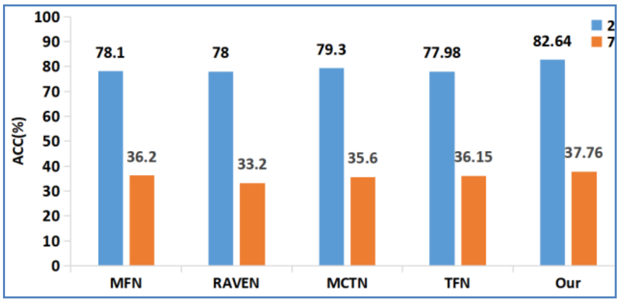

Figure 8. Accuracy comparison of 2&7-classifications on MOSI

As shown in Fig.9, the MAE of all model methods (T+V) and (T+V+A) mode combination is low, which reflects the stability of text and video from the side. It is obvious from the curve that MAE decreases gradually with the increase of modal types, which further verifies that multi-modal emotion classification is superior to single-modal emotion classification.

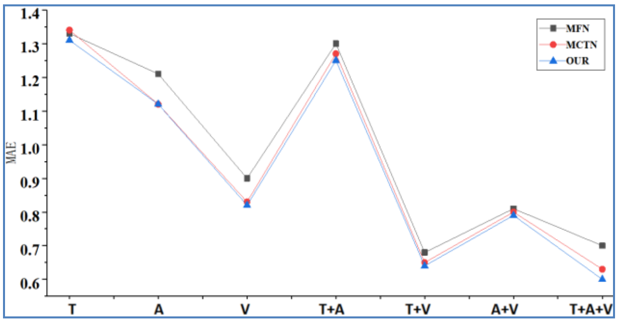

Figure 9. MAE comparison of different models on MOSI

Conclusion

Based on the powerful computing power of computers and the development of deep neural networks, some artificial intelligence systems outperform humans in specific tasks. However, computers still cannot express their emotions through voice, facial expressions, and body movements like human beings. Therefore, how to give computers the ability to recognize and express emotions is a hot research topic. In this paper, based on common multi-modal data such as text, sound, and video, a feature extraction framework of the three-dimensional dynamic expansion is established, and further a multi-modal fusion-matched framework based on spatial and temporal feature enhancement is established. It is used to solve the dynamic correlation problems within and between modes respectively, and then model the short and long-term dynamic correlation information between different modes based on the proposed framework. Experiments performed on MOSI datasets show that the emotion recognition model proposed here in this paper can better utilize the more complex complementary information between different modes. Compared with other multi-modal data fusion models, this framework proposed in this paper has significantly improved emotion recognition rate and accuracy when applied to multi-modal emotion analysis.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University Teknologi Malaysia.

Authors' Contribution Statement

- X. Z. contributed to the conception, the design and implementation of the research, and the writing of the manuscript.
- N. H. M.R. contributed to design of the research， the analysis of the results and interpretation .
- H. H. contributed to data acquisition, data analysis, revision and proofreading of the research.

References

1. Tan, Y., Zhang, J., & Xia, L. A survey of sentiment analysis on social media[J]. Data Anal. Knowl. Discov. .2020;4(1): 1-11. https://doi.org/10.11925/infotech.2096-3467.2019.0769.
2. Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access.2020;8: 168865-168878. https://doi.org/10.1109/ACCESS.2020.3023871.
3. Thandaga Jwalanaiah, S. J., Jeena Jacob, I., & Mandava, A. K. Effective deep learning based multimodal sentiment analysis from unstructured big data. Expert Systems.2023; 40(1): e13096. https://doi.org/10.1111/exsy.13096.
4. Xuyang, W. A. N. G., Shuai, D. O. N. G., & Jie, S. H. I. Multimodal Sentiment Analysis with Composite Hierarchical Fusion. Front. Comput. Sci. .2023; 17(1): 198-208. https://doi.org/10.3778/j.issn.1673-9418.2111004.
5. Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence.2018; (Vol. 32, No. 1):5642-5649. https://doi.org/10.1609/aaai.v32i1.12024.
6. Semma A, Hannad Y, Siddiqi I, et al. Writer identification using deep learning with fast keypoints and harris corner detector[J]. Expert Syst. Appl. . 2021; 184: 115473.https://doi.org/10.1016/j.eswa.2021.115473.
7. Zadeh A, Liang P P, Mazumder N, et al. Memory fusion network for multi-view sequential learning.Proceedings of the AAAI conference on artificial intelligence. 2018; 32(1):5634-5641. https://doi.org/10.1609/aaai.v32i1.12021.
8. Ibrahim, V., Abu Bakar, J., Harun, N. H. ., & Abdulateef , A. F. A Word Cloud Model based on Hate Speech in an Online Social Media Environment[J]. Baghdad Sci. J. 2021;18(2(Suppl.): 0937-0946. https://doi.org/10.21123/bsj.2021.18.2(Suppl.).0937.
9. Hameed, N. H., Alimi, A. M., & Sadiq, A. T. Short Text Semantic Similarity Measurement Approach Based on Semantic Network[J]. Baghdad Sci. J. 2022;19(6(Suppl.):1581-1591. https://dx.doi.org/10.21123/bsj.2022.7255.

10. Gandhi A, Adhvaryu K, Poria S, et al. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Information Fusion. 2022;424-444.https://doi.org/10.1016/j.inffus.2022.09.025.

دراسة تصنيف المشاعر على أساس الاندماج متعدد الوسائط

شيانغ زيهوا[1,2]، نور هيزان محمد رادزي[1]، هاسلينا هاشم[1]

[1]كلية الحاسبات، الجامعة التكنولوجية الماليزية، 81310 جوهور باهرو، جوهور، ماليزيا.
[2]كلية قوانغدونغ للتكنولوجيا، 526100، شارع كيفو، منطقة غاوياو، مدينة تشاوتشينغ، مقاطعة قوانغدونغ، الصين.

الخلاصة

في الوقت الحاضر، لم يعد تعبير الأشخاص على الإنترنت يقتصر على النصوص، خاصة مع ظهور طفرة الفيديو القصير، مما أدى إلى ظهور عدد كبير من البيانات النموذجية مثل النصوص والصور والصوت والفيديو. بالمقارنة مع بيانات الوضع الفردي، تحتوي البيانات متعددة الوسائط دائمًا على معلومات ضخمة. يمكن أن تساعد عملية التنقيب في المعلومات متعددة الوسائط أجهزة الكمبيوتر على فهم الخصائص العاطفية البشرية بشكل أفضل. ومع ذلك، نظرًا لأن البيانات متعددة الوسائط تُظهر ميزات سلسلة زمنية ديناميكية واضحة، فمن الضروري حل مشكلة الارتباط الديناميكي داخل وضع واحد وبين أوضاع مختلفة في نفس مشهد التطبيق أثناء عملية الدمج. لحل هذه المشكلة، في هذا البحث، تم إنشاء إطار استخراج ميزة للتوسع الديناميكي ثلاثي الأبعاد بناءً على البيانات المشتركة متعددة الوسائط، على سبيل المثال الفيديو والصوت والنص. إطار عمل مطابق يعتمد على تحسين الميزات المكانية والزمانية، على التوالي لحل الارتباط الديناميكي داخل الأوضاع وفيما بينها، ومن ثم نمذجة معلومات الارتباط الديناميكي قصيرة وطويلة المدى بين الأوضاع المختلفة بناءً على الإطار المقترح. تُظهر التجارب الجماعية المتعددة التي تم إجراؤها على مجموعات بيانات MOSI

أن نموذج التعرف على المشاعر الذي تم إنشاؤه بناءً على الإطار المقترح هنا في هذه الدراسة يمكنه الاستفادة بشكل أفضل من المعلومات التكميلية الأكثر تعقيدًا بين البيانات المشروطة المختلفة. بالمقارنة مع نماذج دمج البيانات متعددة الوسائط الأخرى، فإن إطار دمج البيانات متعدد الوسائط القائم على الاهتمام المكاني والزماني المقترح في هذه الورقة يحسن بشكل كبير معدل التعرف على المشاعر ودقتها عند تطبيقها على تحليل المشاعر متعدد الوسائط، لذلك فهو أكثر جدوى وفعالية.

الكلمات المفتاحية: مطابقة الميزات، الارتباط الديناميكي، تصنيف المشاعر متعدد الوسائط، مباراة الانصهار، الاهتمام الزمني.