Low-rank Multimodal Fusion Algorithm Based on Context Modeling

Zongwen Bai[1], Xiaohuan Chen[1], Meili Zhou[1], Tingting Yi[1], Wei-Che Chien[2]

[1] School of Physics and Electronic Information, Yan'an University, China
[2] Department of Computer Science and Information Engineering, National Dong Hwa University, Taiwan
ydbzw@yau.edu.cn, GRdh1227@163.com, zml@yau.edu.cn, tt_y3658@163.com, wcc@gms.ndhu.edu.tw

Abstract

As an important part of human daily life, video contains rich emotion information. Therefore, it is a current research trend to find efficient approaches to conducting emotional analysis on videos. Based on tensor fusion, we propose a low-rank multimodal fusion context modeling. At the beginning, modality information is preprocessed by GRU (Gate Recurrent Unit) in Recurrent Neural Network. We construct semantic dependencies to convey contextual information in the context of the video. The proposed model improves performance of applied emotion classification. Additionally, LMF (Low-rank Tensor Multimodal Fusion) with the advantage of end-to-end learning is implemented as a fusion mechanism to improve classification efficiency. We implemented the experiments on CMU-MOSI, POM, and IEMOCAP of multi-modal sentiment analysis, speaker traits and emotion recognition. And results show that our method improved the performance by a margin of 2.9%, 1.3%, and 12.2% respectively contrast with TFN (Tensor Fusion Network).

Keywords: Neural architecture search, Sequence regression models, Performance prediction, Network structure feature

1 Introduction

Multi-modal fusion is an emerging research field of artificial intelligence. Recent works have made great progress in many areas, such as face recognition [1], emotion recognition [2], behavior recognition [3], visual question-answering, and other directions [4], the multi-modal fusion method pushes these technologies into a new stage of development.

Computer applications that use machine learning in multi-modal emotion analysis have become a new trend in the field of fusion between natural language processing and computer vision [5-6]. Qu proposed a dynamic facial expression recognition algorithm based on a deep residual network, which used the thick-trained deep residual network as a feature extractor [7]. Morency proposed a model integrating video, audio and text features, this model can effectively identify emotions from web videos [8]. Su proposed Attention-of-Emoticons Based Convolutional Neural Network (AEB-CNN), which combined emotion and attention mechanism CNN, by adding emoticons to the calculation of attention weight, important keywords get higher attention, thus improving the accuracy of prediction [9]. Bai proposed to use a convolutional neural network and short and long time memory to compress together and accelerate the processing of visual question answering system at the same time, including using tensor to decompose the full connection layer of CNN and LSTM [10]. Recently, text-based sentiment analysis is extended to videos expressing opinions. There are three modalities in the video: language (spoken words), visual (gesture), and acoustic (voice). We focus on the problem of the fusion of different modality features faced by multimodal sentiment analysis.

Due to different modalities information contradict with each other since they are extracted by different features extraction methods, so we proposed a method to fix it, the method combines the context modeling and low-rank tensor network, can model the inter and among model's interaction, and it is end-to-end learnable. Additionally, the proposed is computationally fast and memory saved. Experimental evaluation of multi-modal tasks on three common datasets shows that the proposed model performs better than the previous baseline standard models.

Section 2 of this paper introduces related work of multi-modal sentiment analysis. Section 3 describes the context-modeling based low-rank multi-modal fusion learning method proposed in this paper. Section 4 introduces test parameter setting, test results, and theoretical analysis. Section 5 gives the conclusion of the text and sets future research.

2 Related Work

Previous studies on multi-modal fusion mainly focus

*Corresponding Author: Zongwen Bai; E-mail: ydbzw@yau.edu.cn
DOI: 10.53106/160792642021072204018

on early and late fusion. Early fusion methods simply stack the features of various modalities and fuse these features to the input level [11]. Unfortunately, such approaches cannot fully explore the dynamic effects within the modality. Contrary to early fusion, late fusion enables decisions to be made on the basis of each modality and the decisions to be fused by weighted averaging [12], but cross-modal interaction cannot be modeled since features cannot interact dynamically with each other. Zadeh proposed a memory fusion network, which used LSTM fusion features (Long-Short Term Memory network) over time and extended it by using a dynamic fusion graph to fuse features [13]. Hu proposed an adversarial encoder-decoder-classifier framework to learn a modality-invariant embedding space by introducing adversarial training to match distributions, modality gap can be significantly narrowed and the representations can be directly fused [14].

Recent studies have focused on the dynamic interactions in and among modalities. Arachchi proposed a model composed of Convolutional Neural Networks (CNN), Long Short-term Memory (LSTM), and Gated Recurrent Unit (GRU), by fine-tuning the parameters of the push layer and using the serial LSTM and GRU model [15]. Zadeh proposed a tensor fusion network which creates a tensor representation by calculating the outer product of three different unimodal modalities [16]. That method adopted tensor representation to model interactions among different modalities. However, the outer products of tensors in multi-modal leads to high dimensional, and computational complexity increases exponentially. Mai proposed a locally confined modality fusion network, which contains a bidirectional multi-connected LSTM, as a result local interaction in learning features of each block was improved [17]. Compared with the previous tensor method, the approach contained fewer parameters and training time is shorter, but dividing the feature vector into equal segment influents the performance.

Inspired by the above works, we developed a new model of feature fusion, the modal based on modeling semantic dependencies within each video modality by GRU, adopted a low-rank tensor network to fuse the context-aware multi-modal features, the fusion results are expressed by multi-modal, and softmax function is used to predict the emotion.

3 Proposed Method

According to the relevant methods mentioned above, we proposed a method that combines contextual modeling with low-rank tensor fusion. The model framework involved in the proposed method is shown in Figure 1. The context-aware multi-modal features are obtained by taking the context-modeling multi-modal representation as the input to the sub-embedded network, and then the low-rank tensor fusion network is employed for feature fusion.

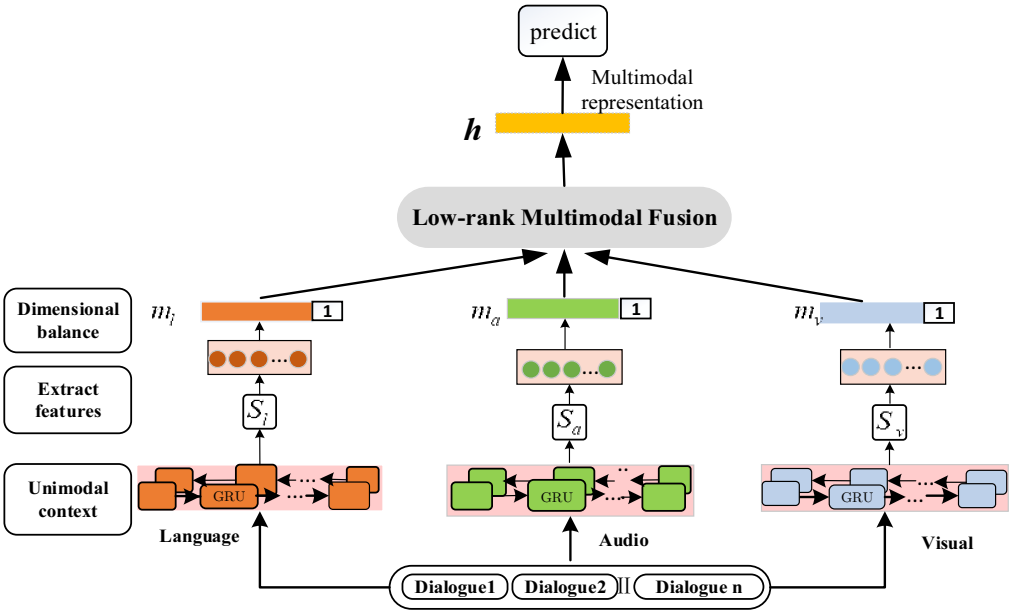

Figure 1. The overall architecture of low-rank multimodal fusion based on context modeling

3.1 Context Modeling

Each modality is interrelated internally in the video, this paper uses context modeling to determine the complete meaning of the utterance in this modality. Based on Zadeh's proposal to use RNN (Recurrent Neural Networks), especially GRU, to model the context dependence in three modalities of video. For each modality in the video session, GRU is used to model the context dependency, and the multi-modal

context dependency representations $l = GRU_l(N_l)$, $a = GRU_a(N_a)$ and $v = GRU_v(N_v)$ are obtained. The following three terms are used to represent unimodal features: $z_a \in \mathbb{R}^{N \times d_a}$ (audio feature) $z_v \in \mathbb{R}^{N \times d_v}$, (visual feature) and $z_l \in \mathbb{R}^{N \times d_l}$ (text feature), where N is the maximum number of utterances in the video.

Therefore, context-aware multimodal features can be expressed as $z_a = S_a(GRU_a(N_a))$, $z_v = S_v(GRU_v(N_v))$, $z_l = S_l(GRU_l(N_l))$.

3.2 Feature Extraction

The dataset contains three modalities (language, visual, acoustic), therefore, we have designed three unimodal embedded networks S_l, S_v, S_a, three sub-embedded networks are used to extract context-aware multimodal representations L, A, V, and obtain context-aware multi-modal features z_l, z_a, z_v and the dimensions of the multimodal features are balanced m_a, m_v, m_l.

In our model, the sub-embedded network of visual and acoustic is a simple 2-layers feed-forward neural network, which is used to extract features. For language modality the sub-embedded network uses LSTM to extract text features.

Language embedded sub-network (S_l) is presented in Figure 2. The Glove is a set of words represented by a sequence of 300-dimension word vectors [18], let language feature be a vector $l = \{l_1, l_2, l_3, ..., l_T; l_t \in \mathbb{R}^{300}\}$, where T_l is the number of words in an utterance. LSTM network is used to learn the time-dependent language representations $L = \{L_1, L_2, L_3, ..., l_T; L_t \in \mathbb{R}^{128}\}$, according to the following LSTM formula.

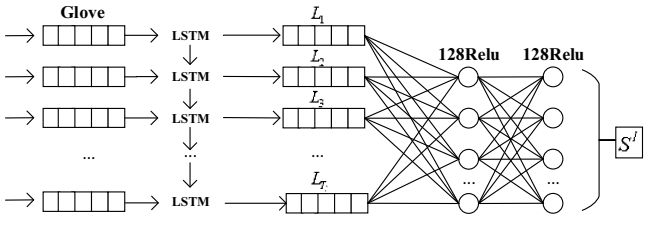

Figure 2. Spoken language embedding subnetwork

$$\begin{pmatrix} i \\ f \\ o \\ m \end{pmatrix} = \begin{pmatrix} sigmoid \\ sigmoid \\ sigmoid \\ tanh \end{pmatrix} W_{l_d} \begin{pmatrix} X_t W_{l_e} \\ L_{t-1} \end{pmatrix} \quad \textbf{(1)}$$

$$c_t = f \odot c_{t-1} + i \odot m \quad \textbf{(2)}$$

$$L_t = o \otimes tanh(c_t) \quad \textbf{(3)}$$

$$L = \left[L_1; L_2; L_3; ...; L_{T_l} \right] \quad \textbf{(4)}$$

Where L is a language representation matrix concatenated by $L_1, L_2, L_3, ..., L_{T_l}$, and L is took as input of fully connected network, and then network generates language embedding $s^l = S_l(L; W_l) \in \mathbb{R}^{128}$, where W_l is collection of weights in S_l network.

Visual embedding sub-network (S_v) is presented in Figure 3. Let visual feature be a vector $\hat{v}_j = [v_j^1, v_j^2, v_j^3, ..., v_j^p]$ of p visual features in j video frame, and T_v represents the total number of video frames. We perform the average pooling on frame to obtain the desired visual feature $v = \left[E\left[v^1\right], E\left[v^2\right], E\left[v^3\right], ..., E\left[v^j\right] \right]$, then V is used as visual embedded network S_v input. We have used FACET extracted information from video, for which deep neural network is used to generate visual embedded features [19]. The network contains three hidden layers of 32 ReLU units with weight matrix W_v. The subnet output provides visual embedding $s^v = S_V(v; W_v) \in \mathbb{R}_{32}$.

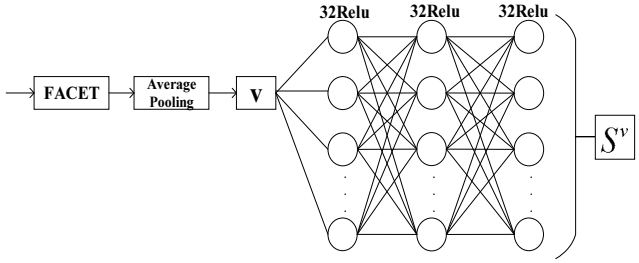

Figure 3. Visual embedding subnetwork

Acoustic embedding sub-network (S_a) is presented in Figure 4. For each opinion fragment with T_a audio frames (sampled at 100Hz; that is 10ms), and extract acoustic feature vector $\hat{a}_j = [a_j^1, a_j^2, a_j^3, ..., a_j^p]$ of p features in j audio frame. These extracted acoustic features are combined on average to obtain desired acoustic feature $a = \left[E\left[a_1\right], E\left[a_2\right], E\left[a_3\right], ..., E\left[q\right] \right]$, where a is input of the audio embedded sub-network S_a. Applied Cov-AREP (Collaborative Speech Analysis Library) can extract rich features from audio, acoustic modality can be modeled using deep neural network [20]. Similar to S_v, S_a is a 3-layers network composed of 32 ReLU units with weight matrix W_a. The subnet output provides acoustic embedding $s^a = S_a(a; W_a) \in \mathbb{R}_{32}$.

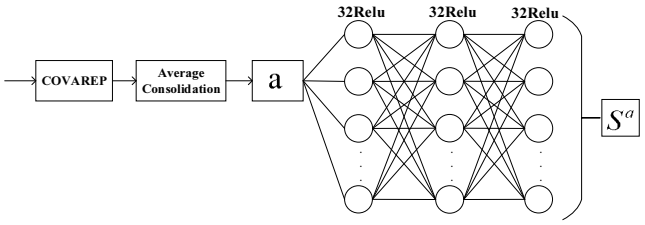

Figure 4. Acoustic embedding subnetwork

3.3 Low-Rank Tensor Multimodal Fusion

The goal of multimodal fusion is to integrate unimodal representation into a compact multi-modal representation for downstream prediction tasks. The tensor is created by taking the outer product on the input modality. In addition, the interactions among the modality subsets are modeled with tensors, and the input tensor M formed by the unimodal representation can be calculated by formula (5).

$$M = \bigotimes_{u}^{U} m_u, m_u \in \mathbb{R}^{d_u} \tag{5}$$

where $\otimes_{u=1}^{U}$ represents the tensor outer product on a set of vectors indexed by u, and m_u is the input representation plus 1. Input tensor $M \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_u}$ generates vector representation through linear layer $g(\cdot)$.

$$h = g(M; W, b) = W \cdot M + b \;\; h, \; b \in \mathbb{R}^{d_y} \tag{6}$$

Where W is the weight of the layer and b is the bias. When M is a tensor of order U (where U is the number of input modalities), the weight of the (U+1) order tensor in $\mathbb{R}^{d_1 \times d_2 \times \cdots \times d_u \times d_h}$ is still W. The dimension of M will increase exponentially with the number of modalities $\prod_{u=1}^{U} d_u$, and the number of parameters learned in the weight tensor W still also increase exponentially, which not only introduces a lot of calculations, but also exposes the model to the risk of overfitting.

In this part, it is explained in detail that LMF decomposes weight W into low-order factors, which reduces the number of parameters in the model. The parallel decomposition of low-rank weight tensors and input tensors is used to calculate tensor-based fusion, which can effectively perform weight decomposition, and the method can scale linearly with a number of modalities.

3.3.1 Low-Level Weight Decomposition

The LMF is to decompose the weighted tensor W into U groups of modality features factors (as shown in Figure 5). Since W is U+1 order tensor, the commonly used decomposition methods will produce U+1 parts. Therefore, W is summed by U-order tensor $\widetilde{W}_k \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_u}$, $W = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} w_m^{(i)}$ and we can decompose each $W = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} w_m^{(i)}$, decomposed into the form of a vector:

$$\widetilde{W}_k = \sum_{i=1}^{r} \bigotimes_{u=1}^{U} w_{u.k}^{(i)}, w_{u.k}^{(i)}, \in \mathbb{R}_u^d \tag{7}$$

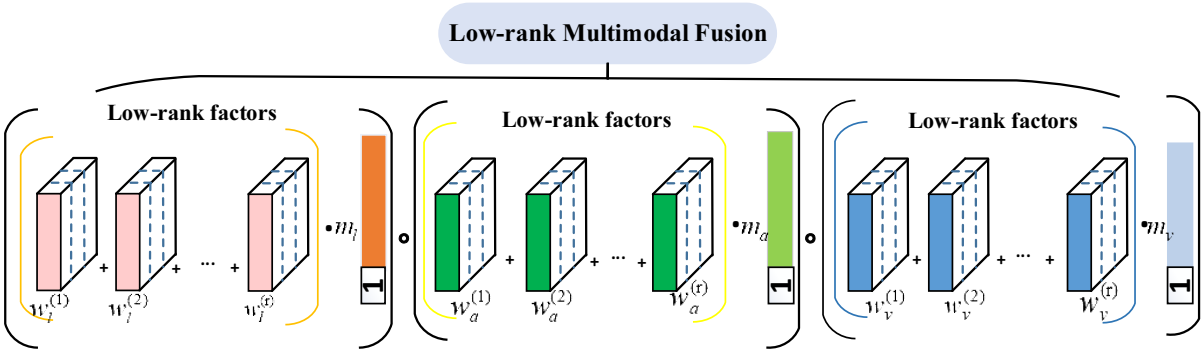

Figure 5. The idea of low-rank multimodal fusion is to decompose the weighted tensor W into a modal specific factor of U

Where the minimum R is called the rank of the tensor that makes decomposition effective, and the vector set $\{\{w_{u,k}^{(i)}\}_{u=1}^{U}\}_{i=1}^{R}$ is called the decomposition factor of the original tensor. In LMF, we start with a fixed rank r and use the decomposition factor $\{\{w_{u,k}^{(i)}\}_{u=1}^{U}\}_{i=1}^{r}$, $W = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} w_m^{(i)}$ to parameterize the model. The decomposition factor can also be used to reconstruct low-order $W = \sum_{i=1}^{r} \bigotimes_{m=1}^{M} w_m^{(i)}$. We can recombine and connect these vectors to the specific low-rank factor of the U group of modalities. Let $w_u^{(i)} = [w_{u,1}^{(i)}, w_{u,2}^{(i)}, ..., w_{u,d_h}^{(i)}]$, for the modality u, $\{w_u^{(i)}\}_{i=1}^{r}$ is its corresponding low-order factor, and the low-order weight tensor is reconstructed by the following method.

$$W = \sum_{i=1}^{r} \bigotimes_{u=1}^{U} w_u^{(i)} \tag{8}$$

Therefore, formula (6) can be calculated by formula (9):

$$h = \left(\sum_{i=1}^{r} \bigotimes_{u=1}^{U} w_u^{(i)} \right) \cdot M \tag{9}$$

Note that for all u, $w_u^{(i)} \in \mathbb{R}^{d_u \times d_h}$ shares the same size in the second dimension, but by introducing a low-rank factor, it is necessary to calculate the reconstruction of $W = \sum_{i=1}^{r} \bigotimes_{u=1}^{U} w_u^{(i)}$ for forward propagation, which generates more computational complexity. Therefore, the following parallel decomposition method is proposed.

3.3.2 Efficient Low-Rank Fusion Using Parallel Decomposition

The original input $\{\mathrm{M}_u\}_{u=1}^{U}$ is decomposed by tensor M, which is parallel to the process of a decomposing tensor into specific modality low-rank factors. Using this principle, formula (9) can be simplified:

$$\begin{aligned} h &= \left(\sum_{i=1}^{r} \bigotimes_{u=1}^{U} w_u^{(i)} \right) \cdot M \\ &= \sum_{i=1}^{r} \left(\bigotimes_{u=1}^{U} w_u^{(i)} \right) \cdot M \\ &= \bigwedge_{u=1}^{U} \left[\bigotimes_{u=1}^{U} w_u^{(i)} \cdot m_u \right] \end{aligned} \tag{10}$$

Where $\bigwedge_{u=1}^{U}$ represents the product of elements on the tensor sequence $\wedge_{t=1}^{3} x_t = x_1 \circ x_2 \circ x_3$.

Figure 1 shows three modalities condition of formula (10), which can be derived from formula (11) in two modalities condition.

$$\begin{aligned} h &= \left(\sum_{i=1}^{r} w_a^{(i)} \otimes w_v^{(i)} \right) \cdot M \\ &= \left(\sum_{i=1}^{r} w_a^{(i)} \otimes m_a \right) \circ \left(\sum_{i=1}^{r} w_v^{(i)} \otimes m_v \right) \end{aligned} \tag{11}$$

An important reason for this simplification is that it utilizes parallel decomposition of M and W, so that h can be calculated without actually creating tensor M from the input representation m_u. In addition, different modalities are decoupled in the simplified calculation of h, which allows the method proposed in this paper to be extended to any number of modalities. Adding a new modality can be done by adding another set of modality-specific factors. Equation (10) is a differentiable operation, and the parameter $\{\mathrm{w}_u^{(i)}\}_{i=1}^{r} u = 1, ..., U$ can be optimized end-to-end through backpropagation.

Using the formula (10), h can be calculated directly from the input unimodal representation and decomposition factor of its modality, avoid the complicated calculation of the input tensor M and weight W. On the contrary, the input tensor and linear projection are implicitly calculated together in formula (10), which makes the tensor fusion method more efficient. In fact, LMF reduces the computational complexity of tensor quantization and fusion from $o\left(d_y \prod_{u=1}^{U} d_u\right)$ to $o\left(d_y \times r \times \prod_{u=1}^{U} d_u\right)$.

In practice, a different form of the formula (10) is used for less complex calculations, formula (12), which connects low-rank factors into a third-order tensor of U, and exchanges the order of element product and summation:

$$h = \bigwedge_{i=1}^{r} \left[\bigwedge_{u=1}^{U} \left[w_u^{(1)}, w_u^{(2)}, ..., w_u^{(r)} \right] \cdot \hat{m}_u \right]_{i,:} \tag{12}$$

The order of summation is backward according to the first dimension of the matrix in parentheses and $[\cdot]_{i,:}$ represents the i-th part of the matrix. In this paper, the U-order tensors are used to parameterize the model instead of vectors.

4 Experiments

The paper adopts tensor fusion network (TFN) as baseline because its structure is the most similar to this paper. The difference is that it is clearly formed as a multi-dimensional tensor for the fusion of different modalities.

4.1 Datasets

Experiments were performed on three multi-modal datasets CMU-MOSI [21], POM [22] and IEMOCAP [23]. The CMU-MOSI dataset is a set of 93 comment videos from YouTube film reviews. The CMU-MOSI dataset has rich emotional expressions and is an annotated dataset. These videos are segmented according to opinions and discourses to adapt to the spoken language whose sentence boundaries are not as clear as the text. Each video consists of multiple opinion segments, and each segment is annotated with a sentiment in the range [-3, 3], where -3 represents highly negative and 3 means highly positive. The POM dataset consists of 903 movie review videos. Each video has traits of the speaker: self-confidence, enthusiasm, pleasant voice, dominant, credible, vivid, professional, entertaining, introverted, trusting, relaxed, extroverted, thorough, nervous, persuasive, humorous. The IEMOCAP dataset contains 151 recorded conversations, each has 2 speakers and a total of 302 videos in the dataset. Each video has 9 emotion annotations: anger, excitement, fear, sadness, surprise, frustration, happiness, disappointment, neutrality. IEMOCAP is the most commonly used dataset in

dialogue emotion recognition. It is of high quality and has the advantage of having multi-modal information.

In order to evaluate the generalization of the model in this article, the data is divided into a training set, validation set and test set, and it is ensured that there is no same speaker from the training set in the test set. The data is divided as shown in Table 1.

Table 1. Speaker independent dataset segmentation, training set, validation set, and training set

Datasets	CMU-MOSI	IEMOCAP	POM
Train	1284	6373	600
Val	229	1775	100
Test	686	1807	203

4.2 Multimodal Data Features and Alignment

Each dataset is composed of three modalities: language, visual and acoustic. In order to achieve the same time alignment between different modalities, this paper uses P2FA [24] to achieve alignment, which can align language, visual and audio at word fine-grained. The visual and acoustic features are calculated by taking the average of their feature values within the word interval. The experiment processes the information in the video as follows.

(1) Language view: The Glove embedding was used to encode a sequence of transcribed words into the sequence of word vectors.

(2) Visual view: Facet library [25] is used to extract a set of visual features of each frame (sampling frequency is 30Hz), including 20 facial action units, 68 facial landmarks, head postures, gaze tracking, and HOG features.

(3) Acoustic view: COVAREP acoustic analysis framework is applied to extract a set of low-level acoustic features.

4.3 Benchmark Model

In multi-modal sentiment analysis, speaker traits recognition, and emotion recognition tasks, the model proposed in this article is compared with the following benchmark models.

Support Vector Machine (SVM) is a widely used non-neural network classifier [26]. It trains series of multi-modal features for classification or regression tasks. We extract unimodal features (Section 3.2) and connect them to form multi-modal features, and then apply the feature vector to SVM for final sentiment classification. Bidirectional Context LSTM (BC-LSTM) performs context-dependent fusion of multi-sequence data, maintaining the latest technology of emotion recognition on the IEMOCAP dataset [27]. Multi-Attention Recurrent Network (MARN) is a model for understanding human communication [28], the model uses a neural network component called MAB(Multi-attention Block) to simulate the interaction between different modalities and store it in LSHTM (Long Short-Term Hybrid Memory).

The recursive multi-level fusion network (RMFN) automatically decomposes multi-modal fusion problems into multiple recursive stages, and in each stage, a subset of multi-modal signals is highlighted and fused with previous representation results [29]. This paper model combines a new multi-level fusion process with a Recurrent Neural Network system to model time and intra-mold the interactions. The Tensor Fusion Network (TFN)creates a multi-dimensional tensor to capture interaction of unimodal, bimodal, and trimodal, so as to realize dynamic interaction of different modalities. Memory Fusion Network (MFN) uses multi-view gate memory units to store the change over time internal information of modalities and interaction information among modalities. The methods of extracting features for BC-LSTM, TFN and MFN are methods involved in processing information in video in section 4.2.

4.4 Evaluation Metrics

Based on the provided tags, different evaluation tasks are performed on different datasets. We applied multi-category classification and regression. Multi-classes classification task is applied to three multimodal datasets, and the regression task is applied to the CMU-MOSI and POM datasets. For binary classification and multi-class classification, F1 score and accuracy (Acc) is used to represent model performance. For regression tasks, Mean Absolute Error (MAE) and Pearson Correlation (Corr) are used to express performance. Except for MAE, the higher the value of other indicators, the better the model performance.

5 Results and Discussion

Experimental results of the classification task of this method on three datasets are shown in Table 2 and Table 3. The boldface indicates the numerical value with the best performance.

As shown in Table 2, the article adopts three representative datasets POM, IEMOCAP, and CMU-MOSI, and randomly extracts modalities from them to form three “text+audio”, “text + video”, and “video + audio” research datasets. Table 2 shows the “video + audio” performance among the three two-modalities fusion. The three-modalities combination “text + video + audio” proposed in this article helped to increase accuracy of classification(ACC) values by 9.0%, 1.8% and 6.7% respectively to the datasets. On the basis of “video + audio” model fusion, we proved the substantial significance of multimodal fusion research.

Table 2. Bimodal fusion and trimodal fusion were compared on the data set for emotional analysis tasks

Datasets	POM		IEMOCAP		CMU-MOSI				
Metrics	MAE	Corr	ACC (%)	F1-score (%)	ACC (%)	MAE	Corr	Acc (%)	F1-score (%)
T + A	0.899	0.109	32.2	81.3	81.9	1.873	0.254	67.8	67.8
T + V	0.855	0.264	33.9	83.4	84.8	1.542	0.498	68.9	70.2
V + A	0.887	0.213	34.0	79.0	85.6	1.673	0.378	70.1	69.4
T + A + V (ours)	**0.794**	**0.382**	**43.0**	**85.9**	**87.4**	**0.916**	**0.670**	**76.8**	**76.7**

As shown in Table 3, compared with TFN, the Corr of this method on POM and CMU-MOSI is increased by 0.261 and 0.037 respectively. Compared to TFN on two tasks of IEMOCAP and CMU-MOSI, the F1-score (%) is increased by 2.7% and 3.3%. Our proposed method reduces MAE performance of two tasks of POM and CMU-MOSI by 0.087 and 0.054 compared to TFN. On three task sets, compared to TFN baseline classification accuracy of ACC(%) has increased by 12.2%, 1.3%, and 2.9% respectively. Acoustic and visual forms in the IEMOCAP dataset are better than CMU-MOSI, thus in the sentiment analysis task (CMU-MOSI) MFN model, the classification accuracy of ACC(%) and F1 score performance are better. On the other hand our proposed method still performs well in performance of MAE and Corr. On three task datasets, compared to other baseline models SVM, MARN, RMFN, BC-LSTM, TFN, classification performance of our proposed model is significantly improved. Using proposed low-rank fusion method based on contextual modeling in multi-modal emotion recognition (IEMOCAP), all emotion recognition (IEMOCAP) results on F1 scores are better than in previous baseline models. In multimodal speaker traits recognition (POM), the three evaluation metrics of our model on POM dataset are improved. In multi-modal sentiment analysis task (CMU-MOSI), performance of MAE and Corr is better than other previous models. Experimental results show that the prediction accuracy of the proposed method in sentiment analysis is significantly improved.

Table 3. Bimodal fusion and trimodal fusion were compared on the data set for emotional analysis tasks

Datasets	POM		IEMOCAP		CMU-MOSI				
Metrics	MAE	Corr	ACC (%)	F1-score (%)	ACC (%)	MAE	Corr	Acc (%)	F1-score (%)
SVM	0.897	0.124	32.7	81.3	83.2	1.864	0.054	50.2	50.1
BC-LSTM	0.889	0.274	34.1	82.5	84.6	1.079	0.581	73.9	73.9
MARN	-	-	39.4	83.1	84.9	1.001	0.629	74.5	74.8
RMFN	0.870	0.376	36.9	84.3	85.9	0.967	0.628	75.1	76.2
MFN	0.802	0.356	39.0	83.2	85.2	0.965	0.632	**77.4**	**77.3**
TFN	0.881	0.121	30.8	83.2	86.1	0.970	0.633	73.9	73.4
Ours	**0.794**	**0.382**	**43.0**	**85.9**	**87.4**	**0.916**	**0.670**	76.8	76.7

6 Conclusion

We explored sentiment analysis based on the low-rank multi-modal fusion method using context modeling. GRU in RNN is used to convey the context information in the video segment, and a low-rank tensor fusion network is applied in the fusion mechanism to improve classification performance on emotion recognition tasks. Through experiments and analysis on three datasets of multimodal sentiment analysis that are more generalization, the method in this paper is significantly better than the baseline model.

In future work, our model will be further developed for other multi-modal applications to verify the robustness of the method and work on processing improvements. In addition, our method can be an optional way for future multi-modal research, which is more efficient on classification tasks, and its tensor representation can save memory and computational costs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61761042, 61871460, 61941112), Key Research and Development Program of Yanan (Grant No. 2017KG-01, 2017WZZ-04-01), Key Research and Development Program of Yulin (CXY-2020-066) and the Natural Science Foundation of Shaanxi Province (Grant No. 2020JM-556).

References

[1] S. Liu, D.-Y. Chen, Z.-F. Chen, C.-H. Ru, M. Pang, Research on Face Recognition Technology Based on ESN Multi Feature Fusion, *Journal of Internet Technology*, Vol. 21, No. 5, pp. 1571-1578, September, 2020.

[2] P. Zhong, D. Wang, C. Miao, EEG-Based Emotion Recognition Using Regularized Graph Neural Networks,

IEEE Transactions on Affective Computing, pp. 1-1, May, 2020.

[3] S. Jiang, Y. Qi, H. Zhang, Z.-W. Bai, X. Lu, P. Wang, D3D: Dual 3-D Convolutional Network for Real-time Action Recognition, *IEEE Transactions on Industrial Informatics*, Vol. 17, No. 7, pp. 4584-4593, July, 2021.

[4] Z.-W. Bai, Y. Li, M.-L. Zhou, D. Li, D. Wang, D. Połap, M. Woźniak, Bilinear Semi-Tensor Product Attention (BSTPA) model for visual question answering, *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom, 2020, pp. 1-8.

[5] O. Boydak, X. Li, E. Alvarez, Guest Editorial: Recent Advances in Specific Applications of Communication, Computer Vision, and Networks, *Journal of Internet Technology*, Vol. 21, No. 3, pp. 861-862, May, 2020.

[6] M.-L. Zhou, Z.-W. Bai, T.-T. Yi, X.-H. Chen, W. Wei, Performance predict method based on neural architecture search, *Journal of Internet Technology*, Vol. 21, No. 2, pp. 385-392, March, 2020.

[7] J.-S. Qu, R.-J. Zhang, Z.-W. Zhang, N. Qiao, J.-S. Pan, Image Sequence Facial Expression Recognition Based on Deep Residual Network, *Journal of Internet Technology*, Vol. 21, No. 6, pp. 1579-1587, November, 2020.

[8] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, L.-P. Morency, Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach, *Proceedings of the 16th International Conference on Multimodal Interaction, ACM*, Istanbul, Turkey, 2014, pp. 50-57.

[9] Y.-J. Su, C.-H. Chen, T.-Y. Chen, C.-C. Cheng, Chinese Microblog Sentiment Analysis by Adding Emoticons to Attention-Based CNN, *Journal of Internet Technology*, Vol. 21, No. 3, pp. 821-829, May, 2020.

[10] Z.-W. Bai, Y. Li, M. Woźniak, M.-L. Zhou, D. Li, DecomVQANet: Decomposing Visual Question Answering Deep Network via tensor decomposition and regression, *Pattern Recognition*, Vol. 110, Article No. 107538, February, 2021.

[11] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional mkl based multimodal emotion recognition and sentiment analysis, *Proceedings of IEEE International Conference on Data Mining (ICDM)*, Barcelona, Spain, 2016, pp. 439-448.

[12] O. Kampman, E. J. Barezi, D. Bertero, P. Fung, Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL short papers)*, Melbourne, Australia, 2018, pp. 606-611.

[13] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018, pp. 5634-5641.

[14] D. Wang, Y. Li, L. Ma, Z.-W. Bai, J. C.-W. Chan, Going Deeper with Densely Connected Convolutional Neural Networks for Multispectral Pansharpening, *Remote Sensing*, Vol. 11, No. 22, November, 2019.

[15] S. P. K. Arachchi, T. K. Shih, C.-Y. Lin, G. Wijayarathna, Deep Learning-Based Firework Video Pattern Classification, *Journal of Internet Technology*, Vol. 20, No. 7, pp. 2033-2042, December, 2019.

[16] A. Zadeh, M.-H. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017, pp. 1103-1114.

[17] Z.-W. Bai, J.-H. Tu, Y.-T. Shi, An Improved Algorithm for the Vertex Cover P-3 problem on Graphs of Bounded Treewidth, *Discrete mathematics and theoretical computer science*, Vol. 21, No. 4, pp. 1-13, November, 2019.

[18] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1532-1543.

[19] Q. Zhu, M.-C. Yeh, K.-T. Cheng, S. Avidan, Fast human detection using a cascade of histograms of oriented gradients, *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, 2006, pp. 1491-1498.

[20] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, Covarep: A Collaborative voice analysis repository for speech technologies, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 960-964.

[21] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages, *IEEE Intelligent Systems*, Vol. 31, No. 6, pp. 82-88, November-December, 2016.

[22] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, L.-P. Morency, Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach, *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI 14)*, Istanbul, Turkey, 2014, pp. 50-57.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, *Journal of Language Resources and Evaluation*, Vol. 42, No. 335, pp. 335-359, December, 2008.

[24] A. Zielonka, A. Sikora, M. Woźniak, W. Wei, Q. Ke, Z.-W. Bai, Intelligent Internet of Things system for smart home optimal convection, *IEEE Transactions on Industrial Informatics*, Vol. 17, No. 6, pp. 4308-4317, June, 2021.

[25] W. Dong, J.-S. Wu, Z.-W. Bai, W.-G Li, W. Qiao, Design of Affinity-aware Encoding by Embedding Graph Centrality for Graph Classification, *Neurocomputing*, Vol. 387, pp. 321-333, April, 2020.

[26] C. Cortes, V. Vapnik, Support-vector networks, *Machine Learning*, Vol. 20, No. 3, pp. 273-297, September, 1995.

[27] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual

question answering and visual grounding, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA, 2016, pp. 457-468.

[28] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, L.-P. Morency, Multi-attention recurrent network for human communication comprehension, *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018, pp. 5642-5649.

[29] P. P. Liang, Z. Liu, A. Zadeh, L.-P. Morency, Multimodal language analysis with recurrent multistage fusion, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium, 2018, pp. 150-161.

Biographies

Zongwen Bai is with the Shaanxi Provincial Key Lab of bigdata of energy and intelligence processing, School of physics and electronic information. He is currently pursuing the Ph. D. degree with the School of Computer Science, Northwestern Polytechnical University, Xi'an. His research interests cover Computer Vision, Nature Language Processing and Deep Learning.

Xiaohuan Chen is currently pursuing a master's degree in computer vision and artificial intelligence at Yan'an University, focusing on the Multimodal Fusion of Information.

Meili Zhou received the M.S. degree in signal and information processing from the yanan university in 2008. She is an associate Professor with the School of physics and electronic information, yanan University. Her Interests cover signal processing, Computer Vision and Image Processing.

Tingting Yi is currently studying for a master's degree at Yan'an University. Her research direction is computer vision, and her main research content is Image Processing.

Wei-Che Chien is currently an Assistant Professor with the Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan. His research interests include Wireless Rechargeable Sensor Networks, 5G Mobile Networks, AIoT, Fog Computing and Cloud Computing.