Two-Level Multimodal Fusion for Sentiment Analysis in Public Security

Jianguo Sun 1, Hanqi Yin 1, Ye Tian 1, Junpeng Wu 1, Linshan Shen 1, and Lei Chen 2

1College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang 150001, China
2College of Engineering and Computing, Georgia Southern University, Statesboro, GA 30458, USA

Correspondence should be addressed to Linshan Shen; shenlinshan@hrbeu.edu.cn

Received 28 December 2020; Accepted 21 May 2021; Published 4 June 2021

Copyright © 2021 Jianguo Sun et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Large amounts of data are widely stored in cyberspace. Not only can they bring much convenience to people’s lives and work, but they can also assist the work in the information security field, such as microexpression recognition and sentiment analysis in the criminal investigation. Thus, it is of great significance to recognize and analyze the sentiment information, which is usually described by different modalities. Due to the correlation among different modalities data, multimodal can provide more comprehensive and robust information than unimodal in data analysis tasks. The complementary information from different modalities can be obtained by multimodal fusion methods. These approaches can process multimodal data through fusion algorithms and ensure the accuracy of the information used for subsequent classification or prediction tasks. In this study, a two-level multimodal fusion (TlMF) method with both data-level and decision-level fusion is proposed to achieve the sentiment analysis task. In the data-level fusion stage, a tensor fusion network is utilized to obtain the text-audio and text-video embeddings by fusing the text with audio and video features, respectively. During the decision-level fusion stage, the soft fusion method is adopted to fuse the classification or prediction results of the upstream classifiers, so that the final classification or prediction results can be as accurate as possible. The proposed method is tested on the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets, and the empirical results and ablation studies confirm the effectiveness of TlMF in capturing useful information from all the test modalities.

1. Introduction

The main way of human communication includes language, visual, and acoustic, which can be generally rendered by speaking or writing, images or videos, and encoded phoneme information, respectively. Therefore, it is incomplete and limited to study the opinion, attitude, and sentiment of an entity only through the language or vision modal. Focusing on this point, multimodal research has been popularized and widely used in many fields of machine learning and artificial intelligence, such as tasks of sentiment analysis [1, 2], emotion recognition [3, 4], behavior recognition [5], video captioning [6], and medical diagnose [7–9]. Besides, in the field of public information security, microexpression recognition [10, 11] is the core process in the lie detection task. However, since the microexpressions are short duration and hard to capture by human and machine, most existing investigations are unsatisfactory. Accordingly, using the information complementarity strategy, i.e., applying the voices and audio modalities of the video as auxiliary information in a microexpression recognition task, can effectively increase the accuracy of lie detection.

Multimodal fusion is the concept that integrates information of multiple modalities by classification or prediction [12] and has become one of the most popular research interests in the field of multimodal machine learning. The main advantages of multimodal fusion can be summarized as follows. First, it can obtain more accurate predictions by observing the phenomenon from different modalities. Secondly, a multimodal system can catch the information that is invisible in individual modalities by multimodal fusion. Third, a multimodal system can still operate when one of the
Fusion in the decision level is a higher level of information fusion, which combines information from different data sources classified individually [27]. Generally, decision fusion for multisources data has the potential to improve classification accuracy, compared with the individual data sources. In previous research studies, decision fusion is widely used in several classification tasks [28]. For example, in the remote sensing image classification task, Mazher et al. [29] proposed a decision fusion method in land cover classification. The authors indicate that the proposed decision fusion method has desired generalization performance and can be also applied to other applications with different sensor data. In the medical diagnosis task, Agarwal and Bedi [30] proposed a fusion method using curvelet and wavelet transform and applied it to the medical diagnosis task that combines the images obtained by computed tomography (CT) scan and magnetic resonance imaging (MRI). The experiment shows that the results of two-level fusion are better than those of one feature-level fusion and decision-level fusion. In the natural resource prediction task, DSS was applied to the problem of sustainable greenhouse management [31] and regarded as an effective strategy to balance the sustainability and the profitability of productions.

Motivated by the above discussion, this study proposes a new modal fusion method named TIMF, which produces unimodal embeddings by using a CNN-BiLSTM neural network and achieves the information fusion of text and audio/video embedding based on tensor fusion and decision fusion stages. The cores of our method include (1) a tensor fusion network used to fuse text data with video and audio data, respectively, and (2) a decision-level fusion strategy, which can fuse the classification results. Then, three datasets of multimodal sentiment analysis and emotion classification, i.e., CMU-MOSI [32], CMU-MOSEI [33], and IEMOCAP [34] are used for experiments to evaluate the effectiveness of the proposed methods.

The rest of this study is organized as follows: Section 2 introduces the proposed model; the experimental setup and the experimental results are given in Section 3, and the conclusion is provided in Section 4.

2. Proposed Method

Figure 1 shows the main diagram of the TIMF model that is established on the stages of data preparation, feature extraction, tensor fusion, and decision fusion. To make this framework clearer, detailed descriptions of some components are given in this section.

2.1. CNN-BiLSTM. A convolution neural network (CNN) is utilized to learn features from each modality after the data preparation stage. In this study, the input sequences of the text, audio, and video feature embeddings are defined as $H_t \in \mathbb{R}^{d_i \times l_i}$, $H_a \in \mathbb{R}^{d_a \times l_a}$, and $H_v \in \mathbb{R}^{d_v \times l_v}$, respectively. To extract the features from these input sequences, a 1D temporary convolution layer is used for the time dimension of each input vector. And the outputs of this convolution layer are denoted as $H_{t1} \in \mathbb{R}^{d_i \times l_i}$, $H_{a1} \in \mathbb{R}^{d_a \times l_a}$, and
and deletion of cell state, respectively. By denoting the output gate, which control
the update, maintenance, and deletion of cell state, respectively; \( \sigma \), \( \phi \), \( \theta \).

The output of the CNN is then processed by the BiLSTM layer, which is extended from LSTM and can learn
the context correlation from time-series data. It should be pointed out that LSTM has two special structures: cell state
context correlation from time-series data. It should be considered. First, by sigmoid activation function, the
output of the CNN is then processed by the BiLSTM layer with a certain probability. Note that the input of the
forget gate includes the hidden state \( H_t \) of the upper sequence and the current sequence data \( x_t \).

The first step is to update the output \( f_t \) of the forget gate, which controls whether to forget the cell state of the upper
layer with a certain probability. Note that the input of the forget gate includes the hidden state \( h_{t-1} \) of the
previous sequence and the current sequence data \( x_t \). Thus, the rule of the output is described by the following
equation:

\[
f_t = \sigma(\theta_f [h_{t-1}, x_t] + b_f),
\]

where \( \theta_f \) and \( b_f \) are the sets of weight matrices and biases vectors in the forget gate, respectively; \( \sigma \) represents
the sigmoid activation function.

The second step is to update the information stored in the current cell state \( C_t \). For this purpose, two parts need to
be considered. First, by sigmoid activation function, the output \( i_t \) of the input gate is generated, which contains the
values to be updated. Then, a tanh layer creates a vector of new candidate values \( a_t \) that could be added to the state, as
shown in the following equations:

\[
\begin{align*}
    i_t &= \sigma(\theta_i [h_{t-1}, x_t] + b_i), \\
    a_t &= \tanh(\theta_a [h_{t-1}, x_t] + b_a),
\end{align*}
\]

where \( \theta_i \) and \( b_i \) mean the sets of weight matrices and biases vectors in the input gate, respectively. Note that once \( f_t, a_t, \) and \( i_t \) are determined, the current cell state \( C_t \) can be updated from the last cell state \( C_{t-1} \) according to

\[
C_t = (C_{t-1} \odot f_t) \oplus i_t \odot a_t,
\]

where \( \odot \) and \( \oplus \) mean the Hadamard product and the concatenation operator, respectively.

Now, we can determine the desired output based on the cell state \( C_t \). Let \( \theta_o \) and \( b_o \) be the sets of weight matrices and biases vectors in the output gate, respectively. The main procedure is composed of two steps shown in equations (5) and (6). First, the output gate is to select the parts to be output from \( C_t \), and the corresponding output is given by \( o_t \). Then, by using the Hadamard product between \( \tanh(C_t) \) and \( o_t \), the desired output can be obtained.

\[
\begin{align*}
    o_t &= \sigma(\theta_o [h_{t-1}, x_t] + b_o), \\
    h_t &= o_t \odot \tanh(C_t),
\end{align*}
\]

Finally, the prediction output of the current sequence is updated in accordance with

\[
y_t = \varphi(\theta_y h_t + b_y),
\]

where \( \theta_y \) and \( b_y \) represent the sets of weight matrices and biases vectors in the predict layer, respectively; \( \varphi \) is the
output activation function, e.g., Softmax in classification task or tanh in prediction task.

So far, the forward propagation of LSTM has been described in detail. It can be seen from equations (1)–(7) that
the LSTM network can only utilize the previous input features of a specific period. This characteristic of LSTM
greatly restricts its application in the sentiment analysis task. In order to simultaneously learn the previous and future
context from sequence data, BiLSTM is considered, which combines two separate hidden LSTM layers with the op-
posite directions of the same input.

The architecture of the BiLSTM network is shown in

\[
\begin{align*}
    &\overrightarrow{h_t} = (\overrightarrow{h_{t+1}}, \overrightarrow{h_{t+2}}, \ldots, \overrightarrow{h_t}), \\
    &\overleftarrow{h_t} = (\overleftarrow{h_t}, \overleftarrow{h_{t+1}}, \\
    &\ldots, \overleftarrow{h_{t+1}}), \\
    &x_t = (x_{t+1}, x_t, \ldots, x_n).\end{align*}
\]

The main procedure of BiLSTM can be described by
Specifically, after the CNN-BiLSTM network, we use the tensor fusion layer to learn interactions between different modalities. The inputs of tensor fusion layer are the outputs from the CNN-BiLSTM network. The calculation can be performed by the following equations:

\[ H_{ta} = [H_{t2} \otimes H_{a2}] \]  \hspace{1cm} (9)

\[ H_{tv} = [H_{t2} \otimes H_{v2}] \]  \hspace{1cm} (10)

This process can be described by Figure 3. The fusion information \( H_{ta} \in \mathbb{R}^{d_t \times d_a} \) and \( H_{tv} \in \mathbb{R}^{d_t \times d_v} \) can be obtained after the tensor fusion layer. According to [16], since the output neurons of tensor fusion are easy to interpret and very meaningful in semantics, it is easy for the subsequent layers of the network to decode the meaningful information.

2.3. Decision Fusion. Fusion at the decision level between the classification results of the text-audio and text-video is considered to further improve the results. According to posterior probabilities, each classifier can yield a scoring matrix as the output of the Softmax layer to characterize the confidence level that the network chooses a specific class as the correct output class. Soft fusion is a method to get the new prediction label by fusing the scoring matrix from classifiers. Since soft fusion can increase accuracy in the minimum and worst case, it has been widely used in the medical diagnose field [9, 35]. In this study, the soft fusion method, which can incorporate the scores from separate fusion modalities, is applied to generate the new prediction label in the sentiment analysis task. The weighted combination of the two fusion modalities scores is shown in the following equation:

\[
S_F(c) = W_{ta} \cdot S_{ta}(c) + W_{tv} \cdot S_{tv}(c),
\]  \hspace{1cm} (11)

where \( W_{ta} \) and \( W_{tv} \) denote the weight of text-audio modal and text-video modal, respectively, and satisfy \( W_{ta} = W_{tv} \) at the initial time. \( S_{ta}(c) \) is the score matrix of the text-audio modal for the prediction of class \( c \), \( S_{tv}(c) \) represents the score matrix of the text-video modal, and \( S_F(c) \) stands for the final classification results.

3. Experiment and Discussion

3.1. Datasets. The proposed method is tested by using three public multimodal datasets: CMU Multimodal Corpus of Sentiment Intensity (CMU-MOSI) [32] and CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [33] datasets for sentiment analysis task and Interactive Emotional Dyadic Motion Capture (IEMOCAP) [34] dataset for emotion classification task. The detailed descriptions of these datasets are given as follows.

CMU-MOSI is built on 93 YouTube movie review videos that are segmented into 3,702 utterance segments including 2,199 opinion video segments. Each segment is annotated by a scale in \((-3, 3)\) to reflect sentiment intensity, where \(-3\) and \(3\) represent the extremely negative and extremely positive sentiments, respectively. These segments are rigorously annotated...
According to the subjective sentiment intensity of the visual features and audio features. Moreover, these segments divided into three parts such as train, validation, and test are 16326, 1871, and 4659 data samples, respectively. The raw data and the distribution of sentiment intensity are shown in Figures 4 and 5, respectively. From Figure 5, we can clearly figure out that the number of segments in different sentiment intensity is relatively average.

CMU-MOSEI is the largest dataset of multimodal sentiment analysis tasks. It consists of 23453 sentence utterance video segments from more than 1000 online YouTube speakers and 250 topics. Each segment video is transcribed and properly punctuated, which can be treated as an individual multimodal example. In this case, train, validation, and test partitions contain 16326, 1871, and 4659 data samples, respectively. Figure 6 shows the part of raw data, and the sentiment intensity distributed in (−3, 3) is given in Figure 7, where −3 and 3 represent the extremely negative and extremely positive sentiments, respectively. It is clear from Figure 7 that most segments are labeled with (0, 1), which implies the sentiment intensity is weakly positive. The uneven distribution of the sentiment intensity corresponds to people’s commenting habits.

IEMOCAP is an acted, multimodal, and multispeaker database, which has a free academic license, long duration, and good emotion label. It contains about 12 hours of 302 videos, in which speakers performed 9 different emotions: angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral. In this study, happy, sad, angry, and neutral emotions are chosen for experiment, and the distribution is shown in Figure 8. These considered videos are divided into 4444 segments with emotion annotations. In this case, train, validation, and test partitions contain 2717, 789, and 938 data samples, respectively.

### 3.2. Multimodal Features

For the above datasets, the unimodal features are extracted from the text, audio, and video data modalities by using global vectors for word representation (Glove) [36], COVAREP [37], and Facet, respectively. Moreover, these unimodal features can also be provided by the CMU-MultimodalSDK.

Text features is extracted by Glove, which is an unsupervised learning algorithm for obtaining vector representations of words. The dimension of each text embedding extracted by Glove is 300 for different inputs in the above datasets.

Audio features are extracted by utilizing COVAREP, which is a collaborative and freely available repository of speech processing algorithms. By using COVAREP, we can obtain low-level acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking, voiced/unvoiced segmenting features, glottal source parameters, and peak slope parameters. Each audio feature is extracted on 25-ms frames with a 5-ms shift, and its dimension is 74 for each dataset.

Video features consist of 35 facial action units extracted from each frame by using Facet, which is widely used for extracting facial expression features such as basic and advanced emotions. Thus, the dimension of each video frame feature is 35 for each dataset.

### 3.3. Performance Metrics

In this section, to show the advantages of the proposed method over the existing methods [16, 17, 19] in multiclass classification and prediction tasks, the following performance metrics are chosen for different datasets.

The CMU-MOSI and CMU-MOSEI datasets are labeled in the range of (−3, 3). According to the suggestion of the authors of the datasets, these labels can be divided into different groups. In this paper, two groups are considered. The labels in the first group are divided into two categories: the negative sentiment and the positive sentiment, which are with the labels in the range of (−3, 0) and (0, 3), respectively. In this case, we choose the binary accuracy (Acc-2) and F-score as the performance metrics on the CMU-MOSI and CMU-MOSEI datasets. The 7-class accuracy (Acc-7) is used as the performance metrics in the other group which have 7 sentiment score classification in (−3, 3). In addition, the mean absolute error (MAE) and the correlation between the prediction results and the ground truth label are chosen in these two datasets.

The IEMOCAP dataset is labeled in 9 different emotions, where happy, sad, angry, and neutral emotions are chosen in this study. For this emotion classification task, F-score and the binary accuracy (Acc-2) are selected as the performance metrics in each emotion.

### 3.4. Training Setup

The proposed method is implemented by open-source PyTorch framework, and it is tested and evaluated on the computer with Intel (R) Xeon (R) Silver 4214 at 2.20 GHz CPU, TITAN RTX GPU with 24 GB memory, and 64 GB computer memory. In addition, the hyperparameter is configured for different datasets, where the learning rate and the batch size vary in the range of...
and \((10^{-3}, 10^{-5})\) and \((16, 128)\), respectively, and the epochs are trained 40–100 times by the model. ReLU and sigmoid are used as the activation functions. Mean square error is used as the loss function, and Adam is used as the optimizer. After optimizing the loss function, we use back propagation to update the parameters in the whole network.

3.5. Comparison between TlMF and the Existing Methods.

In this subsection, the comparison between the proposed method and the existing methods, i.e., the tensor fusion network (TFN) [16], low-rank multimodal fusion (LMF) [17], and multimodal transformer (MuLT) [19], is made based on the experimental results of the sentiment analysis.
It is known from [16] that the TFN combines individual modal’s embeddings via calculating three different outer product subtensors such as unimodal, bimodal, and trimodal, and all subtensors are to be flattened as multimodal embedding vectors. As for LMF, it learns the multimodal embedding based on the similar tensor processing to that of the TFN, but with an additional low-rank factor for reducing computation memory. Different from the TFN and LMF, MuLT focus on interactions between multimodal sequences and latently adapt streams from one modality to another and uses the currently popular transformer model to transform one modal into another for modal fusion. In [19], MuLT has achieved state-of-art results in recent years on the multimodal fusion field. Thus, the comparison with MuLT can further test the superiority of TIMF on the sentiment analysis and the emotion classification. To guarantee the comparison fair, the three methods are tested under the same features and experimental environment.

In teams of the sentiment analysis task, the experiments are performed based on the CMU-MOSI and CMU-MOSEI datasets, and the corresponding comparison results are given in Table 1, respectively. From Table 1, it is seen that TIMF can obtain higher Acc-2, F-score, Acc-7, and Corr and lower MAE than TNF, LMF, and MuLT in the CMU-MOSI dataset. In the CMU-MOSEI dataset, it is also seen that TIMF can obtain higher Acc-2, F-score, Acc-7, and Corr than TNF, LMF, and MuLT. In other words, the proposed method is more effective in achieving accurate prediction than the existing methods under these two datasets. On the other hand, the comparison experimental results on the IEMOCAP datasets are given in Table 2. By observing this table, it is clear that the proposed method leads to higher
Acc-2 and F-score than the existing method in different emotions, which implies that more accurate classification can be achieved by TlMF. From the above discussion, it can be concluded that the proposed method has great superiority on accuracy of prediction and classification over the TFN, LMF methods when it is used for sentiment analysis.

3.6. Ablation Study. In this subsection, two series of ablation studies that compare different TlMF’s variants are performed on the basis of the CMU-MOSI, CMU-MOSEI, and IEMOCAP datasets. The first series is to verify the effectiveness of the data fusion strategy in tensor fusion. For this purpose, we set text, audio, and video as the method’s input separately and neglect the tensor fusion and decision fusion, such that the outputs corresponding to different inputs can be directly obtained from the downstream classifier. In this case, the text-only, audio-only, and video-only variants are denoted by TlMF<sub>t</sub>, TlMF<sub>a</sub>, and TlMF<sub>v</sub>, respectively. In the second series of ablation studies, the influence of the tensor fusion and the decision fusion on the performance of the proposed method is illustrated by the following two cases:

Case 1: set text, audio, and video as the input. Discard the tensor fusion from TlMF and denote this variant as TlMF<sub>ntf</sub>. The effectiveness of the tensor fusion can be confirmed by comparing the performance metrics of TlMF<sub>ntf</sub> and TlMF.

Case 2: set the input as the same as in Case 1 and discard the decision fusion from TlMF. To keep the dimension of the output unchanged, the average-fusion is adopted to fuse the classification results of the upstream classifiers. In this case, the variant is defined as TlMF<sub>avg</sub>, and the effectiveness of the decision fusion can be verified by comparing the performance metrics of TlMF<sub>avg</sub> and TlMF.

Based on the above discussion, the experimental results of the ablation studies are provided in Tables 3 and 4. Specifically, Table 3 depicts the comparison results between TlMF and its variants in the sentiment analysis task when different datasets are utilized. It can be seen from the first three rows of these two subtables that Acc-2 and F-score of the text-only method are larger than those of the audio-only and video-only methods, which implies that the text modal can achieve more accurate sentiment analysis than the audio modal and video modal. The main contribution to this phenomenon is that the intrinsic structure of text is more adapted to emotional expression. Thus, it can represent language information better than the other modalities in the sentiment analysis task. On the other hand, by comparing the results of TlMF<sub>ntf</sub> and TlMF (Case 1), it is clear that with the utilization of the tensor fusion, Acc-2, F-score, Acc-7, and MAE are, respectively, improved by 2.6%, 0.44%, 4.2%, and 4.2% on the CMU-MOSI dataset and by 3.96%, 4.1%, 5.01%, and 8.8% on the CMU-MOSEI dataset, which fully demonstrates the advantage of the tensor fusion in improving prediction accuracy of the sentiment analysis. Moreover, following a similar line to the analysis of Case 1, the effectiveness of the decision fusion can be clearly verified according to the comparison between TlMF<sub>avg</sub> and TlMF (Case 2).

For the emotion classification task, Table 4 provides the comparison results between TlMF and its variants on different emotions. By observing this table, there are two points worth emphasizing. First, it is seen from the first three rows that the text-only method leads to higher Acc-2 and F-score than the other methods for different emotions, that is, the text modal can achieve more accurate emotion classification than the audio modal and video modal. The main contribution to this phenomenon is that the intrinsic structure of text is more adapted to emotional expression. Thus, it can represent language information better than the other modalities in the sentiment analysis task. On the other hand, by comparing the results of TlMF<sub>ntf</sub> in Case 1 and TlMF<sub>avg</sub> in Case 2 with those of TlMF, the effectiveness of the tensor fusion and the decision fusion can be clearly verified.
respectively, due to the improvement of Acc-2 and F-score in different emotions. Thus, it can be concluded that the use of these two fusion strategies can effectively improve the accuracy of the emotion classification.

4. Conclusion

Harmful information such as extreme emotions and potential violence widely exists in multimodal data that are stored in cyberspace. Its analysis are great necessary and urgent in the field of the public information security. To this end, this study has proposed a novel TlMF method that contains both data-level and decision-level fusion. A CNN-BiLSTM neural network has been employed to produce the unimodal embeddings. Based on this, the text-audio and text-video fusion embeddings can be then generated by a tensor fusion network. Furthermore, the soft fusion method has been introduced in the decision-fusion stage of TlMF, such that the classification or prediction results can be as accurate as possible. Finally, the testing results on multimodal sentiment analysis and emotion classification tasks have confirmed that the multimodal features learned from the TlMF model can lead to state-of-the-art performance. This fully shows the effectiveness of the proposed method. Moreover, the availability of the network’s components has been verified through the ablation studies.

Data Availability

The data used to support the findings of this study are found at http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed data/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by Fundamental Research Funds for the Central Universities (3072020CFQ0602, 3072020CF0604, and 3072020CFP0601) and in part by 2019 Industrial Internet Innovation and Development Engineering (KY1060020002 and KY1060020008).

References


