Pattern Recognition 136 (2023) 109259

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis

Di Wang, Xutong Guo, Yumin Tian*, Jinhui Liu*, LiHuo He, Xuemei Luo

School of Computer Science and Technology, Xidian University, Xi'an, 710071, China

ARTICLE INFO

Article history:
Received 25 September 2022
Revised 20 November 2022
Accepted 13 December 2022
Available online 15 December 2022

Keywords:
Multimodal sentiment analysis
Transformer
Text-oriented pairwise cross-modal mappings

ABSTRACT

Multimodal sentiment analysis (MSA), which aims to recognize sentiment expressed by speakers in videos utilizing textual, visual and acoustic cues, has attracted extensive research attention in recent years. However, textual, visual and acoustic modalities often contribute differently to sentiment analysis. In general, text contains more intuitive sentiment-related information and outperforms nonlinguistic modalities in MSA. Seeking a strategy to take advantage of this property to obtain a fusion representation containing more sentiment-related information and simultaneously preserving inter- and intra-modality relationships becomes a significant challenge. To this end, we propose a novel method named Text Enhanced Transformer Fusion Network (TETFN), which learns text-oriented pairwise cross-modal mappings for obtaining effective unified multimodal representations. In particular, it incorporates textual information in learning sentiment-related nonlinguistic representations through text-based multi-head attention. In addition to preserving consistency information by cross-modal mappings, it also retains the differentiated information among modalities through unimodal label prediction. Furthermore, the vision pre-trained model Vision-Transformer is utilized to extract visual features from the original videos to preserve both global and local information of a human face. Extensive experiments on benchmark datasets CMU-MOSI and CMU-MOSEI demonstrate the superior performance of the proposed TETFN over state-of-the-art methods.

1. Introduction

With the rapid development of the social media and short video industries, multimodal data from text, video and audio has grown explosively [1]. At the same time, the wide use of capturing devices in combination with the easiness of their usage, their mobility capabilities and their low cost, made it easy to capture sentiment cues from different users [2], which is identical to human language communication. The three kinds of modalities are semantically related and complement each other in the process of expression. Accordingly, a crucial issue in multimodal sentiment analysis is how to design a multimodal fusion scheme to integrate heterogeneous data efficiently [3], so as to learn multimodal representations that contain more sentiment-related information while preserving consistency and differentiated information for each modality.

In order to obtain efficient fusion representations, earlier work such as CATF-LSTM [4] pays more attention to build novel LSTM structures and attention networks to model contextual information for each modality. With the emergence of the transformer, CTFN [5] and MulT [6] perform mutual encoding among modalities based on the idea of modality translation through the transformer-based structures. Representation learning-based methods MISA [7] and Self-MM [8] simultaneously model the consistent and differentiated information among modalities to improve the accuracy of MSA. However, although these methods achieve performance improvements, they mostly treat each modality equally. Nonetheless, the task-related information is not evenly distributed among modalities. Due to ignoring characteristic that different modalities contribute differently to sentiment analysis, the learned fusion representation lacks sentiment-related information, thus reducing the performance of the model.

For the aforementioned issue, in this paper, we design a novel method named Text Enhanced Transformer Fusion Network (TETFN) for MSA. More specifically, we first preprocess the original videos according to the changes in the speakers' expression and video durations, then utilize vision pre-trained model Vision-Transformer (ViT) [9] to extract visual features from original videos. Further, we use LSTM [10] and Temporal Convolutional Networks [11] to encode contextual information for each modality.

* Corresponding authors.
E-mail addresses: ymtian@mail.xidian.edu.cn (Y. Tian), jhliu@mail.xidian.edu.cn (J. Liu).

https://doi.org/10.1016/j.patcog.2022.109259
0031-3203/

The core of TETFN is text enhanced transformer module, which increases the semantic information of fusion representation to learn the consistency among modalities. The TET module consists of two parts: the text-oriented multi-head attention mechanism and cross-modal transformers. The text-oriented multi-head attention mechanism incorporates textual information into audio and vision modalities. The cross-modal transformers focus on modeling pairwise cross-modal mappings for each modality to obtain information from other two modalities. Meanwhile, we also generate unimodal label for each modality to capture the differentiated information among different modalities.

We evaluate our model by performing the sentiment intensity prediction task in sentiment analysis on two benchmark datasets CMU-MOSI and CMU-MOSEI. Experimental results show that our model outperforms state-of-the-art methods on various metrics. Furthermore, ablation study and case study demonstrate the effectiveness of our model.

The contributions of this work can be summarized as following:

- A Text Enhanced Transformer Fusion Network is proposed, which obtains the consistency among modalities through the text-oriented multi-head attention mechanism and text-guided cross-modal mappings and preserves the differentiated information through unimodal prediction.
- We utilize the vision pre-trained model ViT to preprocess and extract features from the original videos to obtain visual features with global and local information.
- Extensive experimental results on two benchmark datasets demonstrate the superiority of the proposed TETFN over several state-of-the-art methods on Multimodal sentiment analysis task.

2. Related work

2.1. Unimodal sentiment analysis

Unimodal sentiment analysis generally refers to the textual sentiment analysis. Sentiment analysis is an important current research area [12]. Previous work first extracts text features with Bag of Words, Bag of ngrams and Bag of means on word embedding (e.g. using word2vec embedding) [13,14], then utilizes traditional machine learning algorithms such as SVM, Naive Bayes and MaxEnt to classify sentiment polarity [15–17]. With the development of deep learning, CNN [18] and LSTM are gradually applied to sentiment analysis and proved to have superior performance. In recent years, with the emergence of Transformer-based text pre-trained models such as Bert and Roberta [19,20], the development of unimodal sentiment analysis has been promoted.

2.2. Multimodal sentiment analysis

With the development of the new media industry in recent years, the number of short videos has surged. When people express their love for products or their views on things, they are no longer limited to simply using text, but rely on short videos and utilize the human language communication containing text, voice, and facial expressions to express sentiment [21].

Earlier multimodal sentiment analysis is divided into early fusion and late fusion. The early fusion method is to simply concatenate the input of different modalities in the shallow layer of the model, and then input the fused features into a single model to complete feature extraction and prediction. MKL [22] utilizes deep convolutional network, RNN and openAIR to extract visual, textual and acoustic features respectively, then concatenates these features and sends them to the decision layer for classification. MFN [23] uses LSTM to model each modality separately after aligning them in the time dimension. The late fusion first makes decisions according to each modality, and obtains the final prediction result by weighted average of the decision results. Onno K [24] obtains respective prediction scores for each modality after feature extraction and performs weighted calculation on the weight of sentiment corresponding to each modality in decision layer to obtain the final result.

Existing multi-stage fusion methods such as TFN [25] and DCC [26] use tensor fusion to generate local tensors and the global fusion, which reduce redundant information and improve the effect of the model. At the same time, there are also methods using decomposition [27,28]. However, these methods do not explicitly explore the joint embedding space before fusion, which makes the modal gap greatly affect the fusion. RNN-based models including GRU and LSTM have made significant progress in exploiting context-aware information of data [29,30]. Furthermoer, CHFusion [31] employs an RNN-based hierarchical structure to map fine-grained local correlations between modalities. Inspired by the great success of the Seq2Seq model in the field of machine translation, methods based on modal translation such as MCTN [32] and MulT [6] have emerged, which perform information fusion through translation between modalities.

With the transformer structure proposed by Google [33], pre-trained language models have developed rapidly. Methods in increasingly focus on utilizing pre-trained models to obtain more informative linguistic features. Among them, MAG-Bert [34] directly adds the RAVEN [35] module to Bert [19] to obtain the fusion representation with superior performance. MISA [7], Self-mm [8] and BBFN [36] no longer utilize the original GloVe word vector [37], but directly use Bert to extract features from the raw sentences. Our method is based on modality translation. What differs from previous work is that we take text modality as a guide for cross-modal mappings, which enhances the role of text modality in sentiment analysis. Meanwhile, the unimodal subtask retains the differentiated information while learning the multimodal similarity.

2.3. Pre-trained models

Early pre-trained language models such as word2vec and GloVe [37] are context-free and cannot solve the problem of polysemy. Subsequently, the context-based dynamic word vector representation with bidirectional LSTM as the core structure ELMo [38] and pre-trained language models after it are context-dependent. The proposal of transformer, a feature extractor using the Self-Attention mechanism, elevates the effect of pre-trained language models to a new level.

As an unsupervised bidirectional self-encoding model, Bert predicts the masked words according to the context in the pre-training process, which can utilize the contextual information of the predicted words at the same time, so the obtained word embedding has linguistic information and long-term dependence. Since the expression of human sentiment is a long-term changing process [39], features with long-term dependency are more representative of human sentiment tendencies.

Transformer structure has been migrated to the field of computer vision since it can still achieve superior performance in image classification tasks without relying on the convolution. On the premise of retaining the original transformer structure, the 2D image information is converted into 1D sequence information and passed into the transformer. Compared with convolution operations that can simply consider local information, the attention mechanism in transformer can comprehensively consider global feature information. At the same time, some multimodal pre-trained models emerge in an endless stream, such as ViLBERT [40] and VL-BERT [41], which solve various kinds of multimodal downstream tasks after pretrained on large-scale datasets.

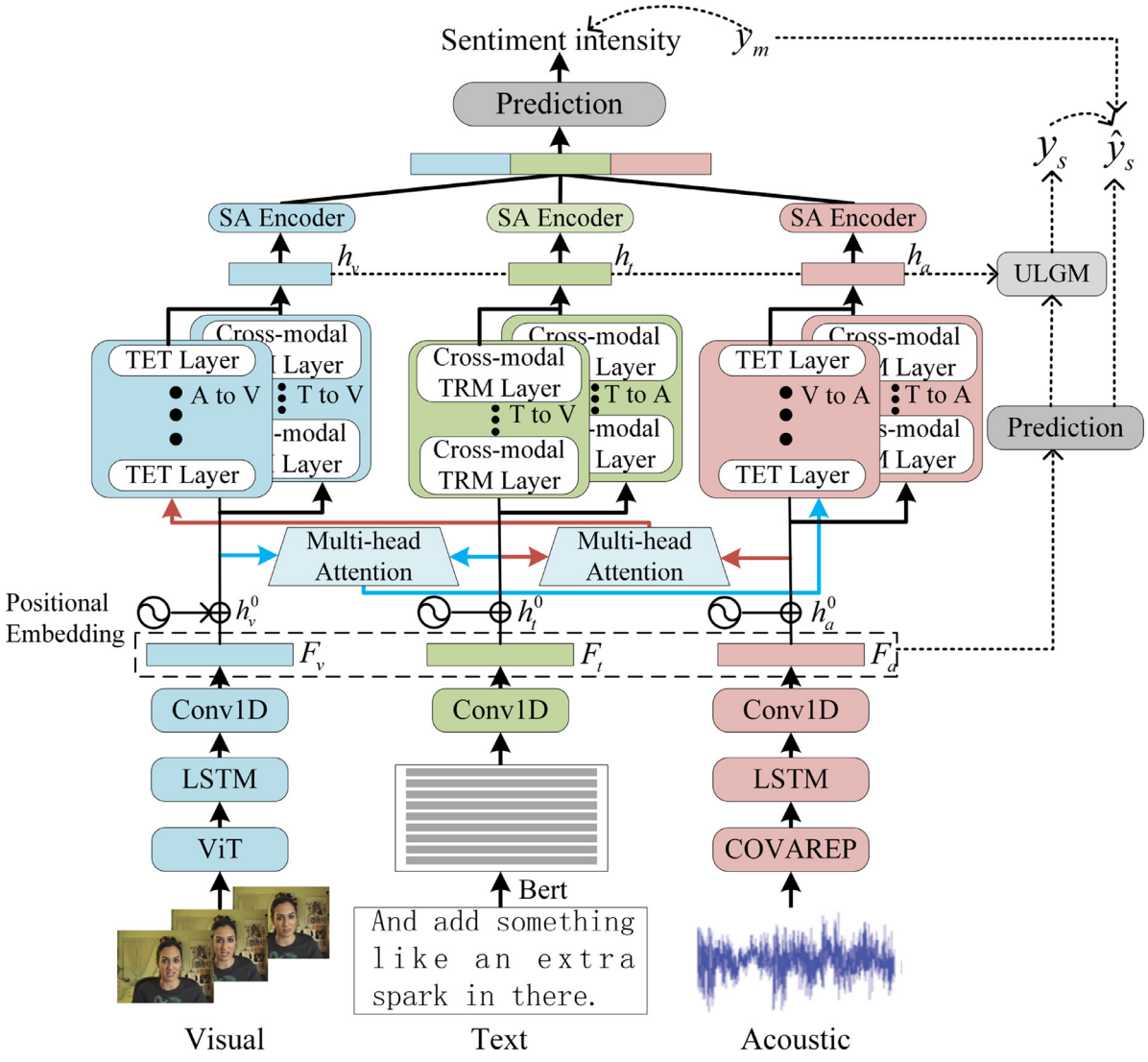

Fig. 1. The overall architecture of TETFN, which is composed primarily of "Feature Extraction and Contextual Encoders", "Text Enhanced Transformer(TET)" and "Unimodal Label Generation Module(ULGM)".

There are two methods for the use of pre-trained models. One is to use the pre-trained models Bert and ViT as language and visual feature extraction modules [7], and the second is to combine visual and acoustic information into the middle layer of Bert [34]. In this paper, we apply the first way to fine-tune Bert to our work, and use ViT to extract visual features from the preprocessed raw videos.

3. Methodology

In this section, we will first define the problem and then describe our TETFN model.

Task Setup. Multimodal sentiment analysis is to judge the sentiment using audio(a), vision(v) and text(t) modalities of the same video fragment, which can be denoted as $I_m \in \mathbb{R}^{T_m \times d_m}$, where T_m represents the sequence length and d_m is the dimension of each modality, $m \in \{a, v, t\}$. We will introduce the way to obtain the feature vectors of text and video in Section 3.2 in detail.

3.1. Overall

Our model is illustrated in Fig. 1. First, features of each modality are extracted from sources and encoded for contextual information, and then the pairwise cross-modality mappings are constructed. In order to strengthen the role of text modality and make cross-modal mappings full of semantic information, for the mappings between vision and audio modalities, the text-oriented multi-head attention is used to make them obtain text-related information. After modeling all modality pairs, the sequence model is utilized for sentiment prediction. Simultaneously, we perform unimodal sentiment analysis module to generate corresponding unimodal labels for each modality.

3.2. Feature extraction and contextual encoder

In order to achieve a superior fusion in the following steps, we first extract textual and visual features with pre-trained models and then encode multimodal sequences before inputting them into the model.

Word Embedding. We utilize the pre-trained language model Bert as the text encoder, which can provide rich semantic information for text modality. Given the raw sentence $S = \{w_1, w_2, \ldots w_n\}$ consisting of n words, after concatenating S with two special tokens [*CLS*] and [*SEP*], the sequence is input into the encoder. Afterwards, the sequence representation with contextual information

$I_t \in \mathbb{R}^{T_t \times d_t}$ is obtained and denoted as the input of text modality.

$$I_t = Bert(S, \theta_{bert}). \tag{1}$$

Visual Feature. As for vision modality, we use a pre-trained vision model Vision-Transformer (ViT) as the vision encoder. The sentiment of vision modality is mainly reflected by facial expressions. Meanwhile, in view of specific organs such as eyes and mouth can better reflect one's sentiment, so we use ViT to obtain both global and local information of the human face.

Given the original video $V = \{p_1, p_2, \ldots, p_n\}$ containing n frames. We assume that the sequence length of the video to be captured is k according to the duration of each video. Only the speaker himself appears in the video, so it is simply necessary to crop the face part while ignoring the complicated background information that has nothing to do with sentiment analysis. Furthermore, we collect frames with open eyes as valid frames, which contain more information related to sentiment, and discard all eyes closed frames. When the number of valid frames is greater than k, we randomly select k frames as the sequence of the video according to the time order of video frames captured and pad the sequence with all valid frames circularly until the sequence length is k if the number of valid frames is less than k. Each valid frame is encoded by ViT and output as a feature vector with a dimension of d_v. After concatenating k vectors, the feature sequence of the video $I_v \in \mathbb{R}^{T_v \times d_v} (T_v = k)$ is obtained as the final feature representation of the vision modality, which is used as the input to the model.

$$I_v = ViT(Prep(V)). \tag{2}$$

Acoustic Feature. For audio modality, we utilize the audio handcrafted features extracted by an acoustic analysis framework named COVERAP [42]. Some of the features include 12 Mel-frequency cepstral coefficients, pitch, volume, glottal source parameters [43], and other features related to emotions and tone of speech. COVERAP feature sequences for every multi-modal example can be obtained by the CMU-MultimodalSDK.

Contextual Encoder. In view of the expression of the speaker's sentiment in the video is a continuous process, the input sequence of each modality is of temporality. Consequently, in order to simulate the changing process of sentiment by injecting temporal long-term dependencies into feature sequences, we use a single-layer long short-term memory network for v and a, followed by a temporal convolutional network for all modalities to capture the time dependence of information at each time step and cast all the hidden states to obtain a uniform dimension for the subsequent process. In this way, we can get:

$$F_v^{lstm} = sLSTM\left(I_v; \theta_v^{lstm}\right), \tag{3}$$

$$F_a^{lstm} = sLSTM\left(I_a; \theta_a^{lstm}\right), \tag{4}$$

$$F_m = \text{Conv1D}\left(\left\{I_t, F_v^{lstm}, F_a^{lstm}\right\}, kernel_m\right), \tag{5}$$

where $F_m \in \mathbb{R}^{T_m \times d}$ and $kernel_m$ represents the size of the convolution kernels of the temporal convolutional network for each modality.

3.3. Text enhanced transformer module

Positional embedding. In order for the model to capture the sequential information of the input, according to the practice of Transformer [33], we add a positional embedding to the low-level representation of each modality.

$$h_m^0 = F_m + PE_m, \tag{6}$$

where $PE_m \in \mathbb{R}^{T_m \times d}$ denotes the positional embedding of each modality, and serving as low-level features for different modalities. h_m^0 is fed into the subsequent network.

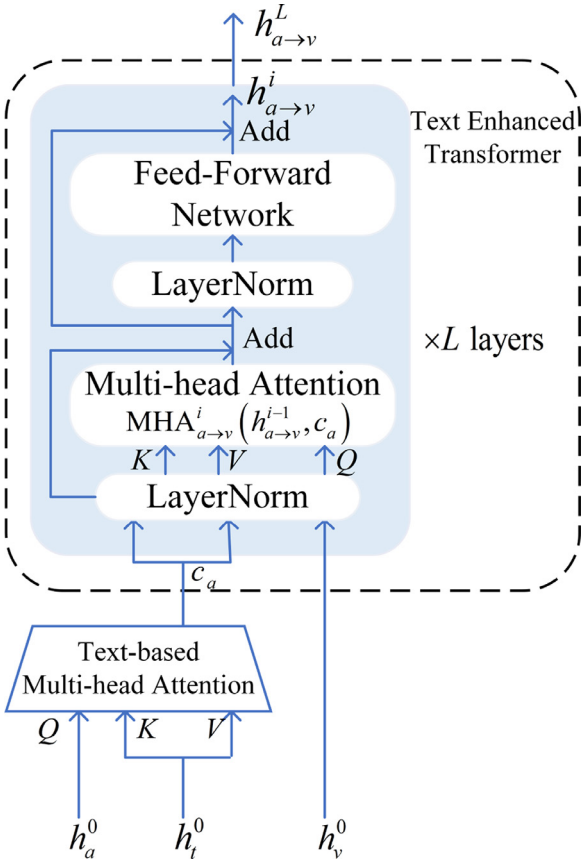

Fig. 2. Text enhanced transformer that models the cross-modal mapping from a to v with text modality enhanced.

Text Enhanced Transformer. According to Mult [6], we design a Text Enhanced Transformer (TET) module to better encode three sources of information (See Fig. 2). TET facilitates one modality to receive information from another by calculating the attention weight between two modalities, thereby promoting the interaction of sentiment-related information for different modalities. As we all know, text is the most basic and intuitive modality that reflects the speaker's sentiment, and contains more sentiment-related information than video and audio. Consequently, when vision modality v and audio modality a are mapped and converted to each other, the fusion representation lacks sentiment-related information and semantics. In response to this situation, besides the standard multi-head attention encoding features from combinations of other modalities, we leverage a text-oriented multi-head attention mechanism, which utilizes text to catalyze the interaction between audio and vision modalities.

We denote the multi-head attention as $output = Multihead(Q, K, V)$. Queries Q, keys K and values V are matrices obtained by projecting representations of three modalities: $Q = h_\alpha^0 W_{Q_\alpha}$, $K = h_\beta^0 W_{K_\beta}$, $V = h_\beta^0 W_{V_\beta}$, where $W_{Q_\alpha} \in \mathbb{R}^{d \times d_f}$, $W_{K_\beta} \in \mathbb{R}^{d \times d_f}$, $W_{V_\beta} \in \mathbb{R}^{d \times d_g}$. The output of the multi-head attention can be obtained by the following equations:

$$output = concat(head_1, head_2, \ldots, head_{h_n}) W^O, \tag{7}$$

$$head_i = Attention(Q, K, V), \tag{8}$$

$$Attention(Q, K, V) = softmax(QK^T \lambda)V, \tag{9}$$

where $Q \in \mathbb{R}^{T_m \times d_f}$, $K \in \mathbb{R}^{T_m \times d_f}$, $V \in \mathbb{R}^{T_m \times d_g}$, $Attention(Q, K, V) \in \mathbb{R}^{T_m \times d_g}$, $W^O \in \mathbb{R}^{h_n d_g \times d}$, h_n is the number of 'head' in multi-head attention.

More specifically, h_a^0 and h_v^0 are used as queries to get the latent adaptation from text to vision and audio:

$$c_v = Multihead\left(h_v^0 W_{Q_v}, h_t^0 W_{K_t}^1, h_t^0 W_{V_t}^1\right), \tag{10}$$

$$c_a = Multihead\left(h_a^0 W_{Q_a}, h_t^0 W_{K_t}^2, h_t^0 W_{V_t}^2\right), \tag{11}$$

where c_v and c_a are intermediate representations of vision modality and audio modality receiving information from text modality. Then c_v and c_a are inputted into transformer encoders and utilized to generate new key-value pairs with original representation of audio and vision modality, respectively.

In order to get the mapping from one modality to another, taking audio to vision as an example ($a \to v$), we stack L layers of text enhanced transformer layer (See Fig. 2), for $i = 1, 2, \ldots, L$ layers:

$$h_{a\to v}^0 = h_v^0, \tag{12}$$

$$h_{a\to v}^i = \mathrm{MHA}_{a\to v}^i\left(\mathrm{LN}\left(h_{a\to v}^{i-1}\right), c_a\right) + \mathrm{LN}\left(h_{a\to v}^{i-1}\right), \tag{13}$$

$$h_{a\to v}^i = \mathrm{FFN}\left(\mathrm{LN}\left(h_{a\to v}^i\right)\right) + \mathrm{LN}\left(h_{a\to v}^i\right), \tag{14}$$

where $h_{a\to v}^0 \in \mathbb{R}^{T_m \times d}$, $h_{a\to v}^i \in \mathbb{R}^{T_m \times d_g}$, the output of each cross-modal transformer layer is used as the input of the next layer. $\mathrm{MHA}_{a\to v}^i$ is the multi-head attention from audio to vision at the i^{th} layer. $\mathrm{LN}(\cdot)$ represents the layer normalization, and $\mathrm{FFN}(\cdot)$ means the feed-forward network.

Identically, after concatenating the output of the last layer of two cross-modal transformers for each modality, the representation of each modality is obtained, which is inputted into a transformer encoder[33] and encoded to acquire self-attention. Ultimately, representations of all modalities are concatenated and passed through fully-connected layers to make sentiment prediction:

$$h_v = \left[h_{a\to v}^N; h_{t\to v}^N\right], \tag{15}$$

$$pred = FC(SA(h_v); SA(h_t); SA(h_a)). \tag{16}$$

Note that what our text enhanced transformer differs from the cross-modal transformer in Mult [6] is that cross-modal transformer treats each modality equally, while TET takes advantage of the characteristic that text modality contributes the most to sentiment analysis among all modalities, which obtains multimodal representation with more sentiment-related information by strengthening the role of text and incorporating textual information into a and v.

Unimodal Label Generation Module. We integrate the unimodal label generation module(ULGM) [8] into our method to capture modality-specific information.During forward propagation, the last hidden-states through LSTM of audio and vision modality are adopted as initial representations. Meanwhile, the first word vector in the last layer of Bert is selected as the textual representation. Then, the unimodal prediction $\hat{y}_s$ is obtained by fully-connected layers. In the training phase, we first define positive and negative centers with predicted unimodal labels and multimodal fusion representations. Afterwards, we calculate the relative distance from the representation of each modality to the positive and negative centers, and obtain the offset value from the uni-modal label to the multi-modal label to generate the uni-modal label $y_s^{(i)}$ for i^{th} epoch. In this way, it is more conducive to sentiment analysis to obtain differentiated information of different modalities while retaining the consistency of each modality.

3.4. Training loss

Finally, combining the two tasks, we use Mean Absolute Error (MAE) as the basis to calculate the loss function for the multimodal task and the unimodal task.

$$Loss_{mul} = \frac{1}{N}\sum_i^N \left(\left|pred^i - y^i\right|\right), \tag{17}$$

$$Loss_s = \frac{1}{N}\sum_i^N \left(\sum_s^{\{t,a,v\}} W^i * \left|\hat{y}_s^i - y_s^{(i)}\right|\right), \tag{18}$$

$$Loss = Loss_{mul} + Loss_s, \tag{19}$$

where N denotes the number of training samples, y represents the true label for multimodal data and $W^i = \tanh\left(\left|y_s^{(i)} - y\right|\right)$ is the weight of i^th sample.

Table 1
Dataset statistics in MOSI and MOSEI.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16,326	1871	4659	22,856

4. Experiments

In this section, we will introduce the datasets, baselines and the experimental settings.

4.1. Datasets

Our experiments are performed on two benchmark multimodal sentiment analysis datasets: CMU-MOSI [44] and CMU-MOSEI [45], the composition of which are shown in Table 1.

CMU-MOSI.CMU-MOSI is a collection of commentary videos. Each video clip is manually labeled with [-3, 3] according to the sentiment intensity expressed by the speaker, representing sentiment ranging from strong negative to strong positive.

CMU-MOSEI. As an upgraded version of CMU-MOSI dataset, it is currently the largest multimodal sentiment analysis dataset. CMU-MOSEI dataset contains 23,353 labeled video clips segmented from 2928 videos, which includes sentiment intensity labels as well as six emotion tags such as happy, sad, etc.

4.2. Baselines

In order to fully guarantee the effectiveness of TETFN, we conduct a fair comparison among the baselines and the state-of-the-art methods in multimodal sentiment analysis.

Unimodal. For three kinds of unimodal sentiment analysis, a single-layer undirectional LSTM is used as encoder followed by a fully-connected layer for prediction.

TFN. The Tensor Fusion Network (TFN) [25] uses modality embedding subnetwork and tensor fusion to learn intra- and inter-modality dynamics.

LMF. The Low-rank Multimodal Fusion (LMF) [46] performs multimodal fusion using low-rank tensors to improve efficiency.

MFN. The Memory Fusion Network (MFN) [23] explicitly accounts for both interactions in a neural architecture and continuously models them through time.

RAVEN. The Recurrent Attended Variation Embedding Network (RAVEN) [35] models the fine-grained structure of nonverbal subword sequences and dynamically shifts word representations based on nonverbal cues.

MULT. The Multimodal Transformer (MulT) [6] uses cross-modal transformer based on cross-modal attention to make modality translation.

ICCN. The Interaction Canonical Correlation Network (ICCN) [47] learns correlations between all three modes via deep canonical correlation analysis (DCCA).

Table 2
Results on CMU-MOSI and CMU-MOSEI. (B) means textual features are based on Bert; [1] means the results provided by Self-MM [8] and [2] is from MAG-Bert [34]. Models with * are reproduced under the same conditions. The left of the "/" is calculated as "negative/non-negative" and the right is calculated as "negative/positive" in F1-score and Acc-2.

	CMU-MOSI				CMU-MOSEI				
Model	MAE	Corr	Acc-2	F1-score	MAE	Corr	Acc-2	F1-score	Data State
A	1.183	0.397	47.83/50.0	45.72/47.47	0.627	0.149	68.63/42.86	55.86/35.71	Unaligned
V	1.092	0.376	71.42/71.42	71.43/71.43	0.619	0.353	68.63/55.86	52.86/45.71	Unaligned
T	0.804	0.698	78.16/80.48	78.19/80.46	0.554	0.490	74.51/75.0	74.71/74.84	Unaligned
TFN(B)[1]	0.901	0.698	-/80.2	-/80.7	0.593	0.700	-/82.5	-/82.1	Unaligned
LMF(B)[1]	0.917	0.695	-/82.5	-/82.4	0.623	0.677	-/82.0	-/82.1	Unaligned
MFM(B)[1]	0.877	0.706	-/81.7	-/81.6	0.568	0.717	-/84.4	-/84.3	Aligned
RAVEN[1]	0.915	0.691	78.0/-	76.6/-	0.614	0.662	79.1/-	79.5/-	Aligned
MulT(B)[1]	0.861	0.711	81.5/84.1	80.6/83.9	0.58	0.703	-/82.5	-/82.3	Aligned
ICCN	0.862	0.714	-/83.0	-/83.0	0.565	0.713	-/84.2	-/84.2	Unaligned
MAG-Bert(B)[2]	0.712	0.796	84.2/86.1	84.1/86.0	-	-	84.7/-	84.5/-	Aligned
MISA(B)[1]	0.783	0.761	81.8/83.4	81.7/83.6	0.555	0.756	83.6/85.5	83.3/85.3	Aligned
Self-MM(B)[1]	0.713	0.798	84.0/85.98	84.42/85.95	0.530	0.765	82.82/85.17	82.53/85.30	Unaligned
MAG-Bert(B)*	0.734	0.789	82.42/84.15	82.45/84.13	0.555	0.758	82.38/85.16	82.07/85.24	Aligned
Self-MM(B)*	0.72	0.792	83.15/84.82	83.12/84.84	**0.533**	**0.761**	81.33/84.63	81.77/84.59	Unaligned
TETFN*	**0.717**	**0.800**	**84.05/86.10**	**83.83/86.07**	0.551	0.748	**84.25/85.18**	**84.18/85.27**	Unaligned

MAG-Bert. The Multimodal Adaptation Gate for Bert (MAG-Bert) [34] incorporates aligned nonverbal information to text representation within Bert.

MISA. The Modality-Invariant and -Specific Representations (MISA) [7] projects representations into modality-sprcific and modality-invariant spaces and learns distributional similarity, orthogonal loss, reconstruction loss and task prediction loss

Self-MM. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning (Self-MM) [8] designs a multi- and a uni- task to learn inter-modal consistency and intra-modal specificity.

4.3. Experimental settings

Evaluation Metrics. We establish two evaluation tasks, regression and classification task, respectively. As for regression task, Mean Absolute Error (MAE) and Pearson Correlation Coefficient (Corr) are reported. For classification task, we use F1-score and Binary accuracy (Acc-2) as evaluation metrics. We calculate Acc-2 and F1-Score in two ways: negative/non-negative (non-exclude zero) [25] and negative/positive (exclude zero) [6]. Except for MAE, higher values are better.

Experimental Details. Our model is trained with Adam optimizer. The initial learning rate of Bert is $5e-5$, and learning rate for other parameters is $1e-3$. The fine-tune range of mini-batch is $\{16, 32, 64\}$ and the fine-tune range of the hidden size of LSTM is $\{32, 64, 128\}$, and the kernel size of Conv1D is 3. We set the sequence length of visual features k is 50. The number of heads for text-oriented multi-head attention is set to 5.

5. Results and analysis

5.1. Quantitative results

The comparative results for multimodal sentiment analysis on CMU-MOSI and CMU-MOSEI are presented in Table 2. According to the different types of datasets, it can be divided into aligned and unaligned. In general, models utilizing aligned datasets will perform better. In this paper, we perform unaligned datasets on our model. As can be seen from Table 2, first, for unimodal baselines A, V, and T, text outperforms nonverbal modalities, which can be observed that people prefer to express their emotion in language. Second, as for multimodal baselines, for the unaligned methods (TFN and LMF), our method outperforms these models greatly. For aligned models (RAVEN and Mult), our method also shows a competitive performance. Also, we reproduce the two best baselines 'Mag-Bert' and 'Self-MM' with the public source codes following the same parameter settings in papers under the same experimental hardware equipment as the proposed method use, respectively. Furthermore, we run the proposed method and two state-of-the-art methods five times under different random number seeds and take the average values as the final results.

5.2. Feature extraction analysis

As for the visual features, we extract high-level features by the pre-trained model ViT instead of the Facet Toolkit which is commonly used by existing methods. In order to extract visual features containing more sentiment-related information, we adopt three kinds of data preprocessing methods on the original videos before sending sampled images into ViT: In the case of no face clipping, equal interval sampling and equal number sampling are performed, respectively. More specifically, the sequence length in the equal interval sampling method is the same as that in using Facet Toolkit, and 10 frames are taken per video for equal number sampling method. The sampling method in the case of clipping the face is described in Section 3.2 in detail. We utilize the same network structure to conduct vision modal sentiment analysis experiments with the features obtained by the above three sampling methods and the features extracted by the original Facet Toolkit, so as to determine the final feature extraction method according to the experimental results.

After sending the preprocessed frame sequence into the ViT, the extracted visual features are passed through the unimodal sentiment analysis network for sentiment analysis to obtain classification, which consists of a single-layer unidirectional LSTM, the ReLU function as the activate function and a fully connected layer as the prediction layer. Furthermore, the hidden state of the last time step of LSTM is obtained as the representation of vision modality. Taking the feature extraction method finally adopted in Section 3.2 as an example, the network structure is illustrated as Fig. 3.

The experimental results are presented in Table 3, from which we can conclude that compared with features extracted by Facet Toolkit, features extracted by ViT perform a higher accuracy of binary classification. For different methods mentioned above, in the case of not clipping the face part, the difference between two sampling methods is slight. However, the result when clipping the face is obviously higher, which means that after removing com-

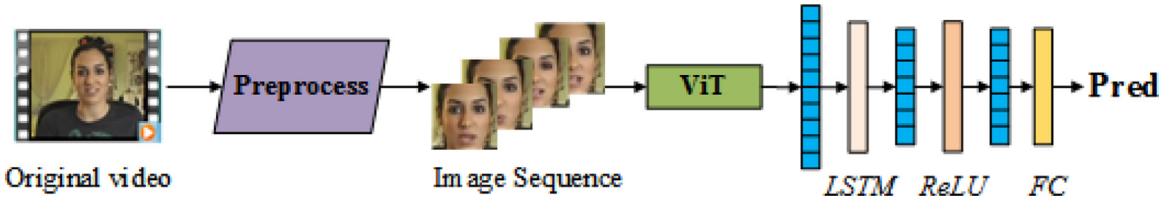

Fig. 3. The network structure for vision modality sentiment analysis.

Table 3
The experimental results of different sampling methods for visual features.

	Acc_2	F1_score
Facet Toolkit	52.17/54.54	51.72/53.98
Vit_with_len	57.41/57.41	55.78/55.78
Vit_10	57.12/57.12	55.36/55.35
Vit_face_selected_50	**71.42/71.42**	**71.43/71.43**

Table 4
Ablation studies on CMU-MOSI dataset. The complete TETFN works best.

	MAE	Corr	Acc-2	F1-score
TETFN	0.717	0.800	84.05/86.10	83.83/86.07
TETFN w/o ViT-feature	0.719	0.792	83.67/85.52	83.59/85.49
TETFN w/o TET	0.733	0.790	82.65/84.76	82.83/84.86
TETFN w/o ULGM	0.970	0.663	79.88/81.40	79.90/81.36
TETFN w/o tMHA	0.723	0.796	83.82/85.67	83.78/85.68

plex background information and keeping eyes open, ViT can focus more on capturing the global and local information of the face part. As a consequence, we finally choose the method described in Section 3.2 to extract features for vision modality. Particularly, what is worth noting is that since vision modality plays the role of auxiliary compared with text modality in multimodal sentiment analysis, so that the accuracy is lower than that of text modality, and the experimental data has certain instability.

5.3. Ablation studies

TETFN consists of three main components: visual features extracted by ViT, text enhanced transformer module and unimodal label generation module.

First, in order to verify the necessity of three components in multimodal sentiment analysis modeling, we conduct several comprehensive ablation studies using the unaligned version of CMU-MOSI to examine the impact of each component. Based on the complete model, we gradually remove the different components. Each version of the model is explained as follows:

TETFN: Our proposed model to enhance the role of text modality in multimodal sentiment analysis.

TETFN w/o ViT-feature: Our model without visual features extracted by ViT. In this case visual features lose both global and local information.

TETFN w/o TET: Our model without text enhanced transformer module. The visual, acoustic, and textual features are simply concatenated and then classified by fully-connected layers, which not only treats all modalities equally, but also does not learn the interaction between different modalities, so it cannot be guaranteed that the final fusion representation contains enough sentiment-related information.

TETFN w/o ULGM: Our model without unimodal label generation module. This will result in the model only learning the consistency but lacking the differentiated information among modalities.

TETFN w/o tMHA: Our model without text-based multi-head attention. This will result in audio and visual representations lacking semantic and sentiment-related information.

Table 4 demonstrates the results of ablation studies using several different variants of our model. The results indicate that the TET module and the ULGM are necessary for achieving state-of-the-art performance. The TET module does model dynamics among three modalities. Visual features with global and local information can improve the expression of sentiment-related information. Finally, it can be observed that text-based multi-head attention injects more semantic information into audio and vision modalities while ensuring the interaction between different modalities.

Second, in order to further verify the contribution of different modalities to multimodal sentiment analysis, in addition to the above-mentioned experiments on the complete model of the three target modalities, we perform experiments of one target modality and two target modalities on CMU-MOSI dataset, respectively. The experimental results are illustrated in Table 5.

In the case of one target modality, we take a, v, t as target modality, and utilize one single cross-modal transformer to obtain cross-modal mappings: $[v, t \rightarrow a]$, $[a, v \rightarrow t]$, and $[a, t \rightarrow v]$. Simultaneously, as for the case of two target modalities, bi-cross-modal transformers are used to model cross-modal mappings: $[v, t \rightarrow a]$ and $[a, v \rightarrow t]$, $[v, t \rightarrow a]$ and $[a, t \rightarrow v]$, $[a, v \rightarrow t]$ and $[a, t \rightarrow v]$. Combined with the experimental results in Table 5, it can be concluded that when there is only one target modality, the highest accuracy can be achieved with text as the target modality; When there are two target modalities, the combination containing t such as $v + t$ and $a + t$ always achieve better performance than combination without t like $a + v$. Under these conditions, the statistic leads us to the conclusion that text contributes more to multimodal sentiment analysis compared with audio and vision modalities.

To sum up, according to the experimental results of variant numbers of target modalities in Table 5, it can be concluded that three target modalities achieve better performance. The conclusion can be drawn that under the premise of strengthening the role of text modality, the information from all modalities can complement each other, which not only promotes the expression of speaker's sentiment in the video, but also obtains higher accuracy for sentiment analysis.

5.4. Case study

In order to prove that highlighting the role of text modality is beneficial to sentiment analysis, we compare the performances when enhancing a (Audio-Enhanced) and v (Vision-Enhanced). For specific case, the results of enhancing different modalities are shown in Table 6. Among them, 'Example' contains the descriptions of different modalities for each case. The column 'Label' represents the real label value of the current video clip with a value between [-3,3], and the real numbers represent the true labels and the predicted values obtained by the regression task when enhancing different modalities.

Combining the three cases presented in Table 6, it can be concluded that strengthening the role of text can improve the perfor-

Table 5
The experimental results of different numbers of target modalities on CMU-MOSI dataset.

Number of target modalities		Acc_2	F1_score	MAE	Corr
One target modality (ours)	$[v, t \to a]$	82.65/84.91	82.58/84.90	0.7430	0.781
	$[a, v \to t]$	**83.09/84.91**	**83.05/84.92**	**0.733**	**0.786**
	$[a, t \to v]$	82.94/84.76	82.92/84.79	0.7724	0.7794
Two target modalities (ours)	$[v, t \to a]$ $[a, v \to t]$	83.67/85.37	83.63/85.37	**0.7116**	0.794
	$[v, t \to a]$ $[a, t \to v]$	83.24/85.06	83.19/85.07	0.7406	0.7873
	$[a, v \to t]$ $[a, t \to v]$	**83.82/85.98**	**83.73/85.95**	0.7156	**0.7913**
Three target modalities (ours)	$[a, v \to t]$ $[a, t \to v]$ $[v, t \to a]$	**84.05/86.10**	**83.83/86.07**	0.717	**0.800**

Table 6
Examples of TETFN, Vision Enhanced(VE) and Audio Enhanced(AE) Fusion Networks.

	Example	Label	TFTFN	VE	AE
T	I love the name Wade.				
V	nod, eyes wide open	1.8000	1.7026	2.2146	1.6409
A	steady and slow				
T	The third one was a piece of crap.				
V	raise eyebrows and roll eyes				
A	lower the volume	−2.400	−2.4857	−2.1357	−1.8992
T	Let me see if I can rustle up some symphaty. Yeah I've got nothing.				
V	frowning, puzzled expression, shaking head	-2.200	−2.2295	−1.4783	−0.9016
A	disgustedtone, sigh				

Table 7
The number of parameters and FLOPs of SOTA methods and TETFN.

Methods	Number of Parameters/M	FLOPs/M	MOSI-Acc_2
Mag-Bert	86.947	17.256	82.48/84.02
Self-MM	85.854	17.011	83.15/84.82
TETFN	87.136	34.531	84.05/86.1

mance of the model in sentiment analysis most. As can be seen from cases 1 and 2, when there are nouns and verbs with more sentiment-related information, the superiority of the model with text-enhanced in predicting sentiment intensity labels is more pronounced. Case 1 and case 3 illustrate that when the sentiment intensity is weak, a smooth tone can better reflect the real sentiment intensity than an exaggerated expression. While if the sentiment intensity is strong, rich facial expressions can better reflect real sentiment than flat intonation.

5.5. Model complexity

In order to further analyse the proposed model in other aspects but not just accuracy, we calculate the FLOPs and the number of parameters for the proposed method and two state-of-the-art methods as shown in Table 7. It can be observed that the parameters of the three methods are basically the same, which indicates that the model complexity differs very limited. In particular, the gap on FLOPs is caused by the high dimension of the visual modality of our method, which is also acceptable.

6. Conclusion

In this paper, we presented the Text Enhanced Transformer Fusion Network (TETFN). TETFN learns the consistency and differentiated information among modalities by modeling text-oriented cross-modal mappings and generating unimodal labels for each modality, of which visual features are extracted by ViT. Compared with previous work, TETFN achieves competitive results in MSA tasks on two benchmark datasets, which strongly proves that strengthening the role of text modality in multimodal sentiment analysis is effective for improving the accuracy of sentiment analysis. Further, we demonstrate the effect of ViT-extracted features, the text enhanced transformer module and the unimodal label generation module by designing multiple ablation studies, respectively. For ViT-extracted features, we conduct several experiments with different sampling methods to choose the most suitable feature extraction method. We also verify the effect of the model under different number of target modalities. Finally, the effectiveness of enhancing text modality of our model is proved through the case study. In future work, we will focus on simplifying the multimodal fusion method and removing redundant information from the fusion representation, thereby further improving the accuracy of multimodal sentiment analysis tasks. We believe this work can inspire the creativity in multimodal sentiment analysis in the future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62072354, 61972302, 62276203 and 62072355, in part by the Key Research and Development Program of Shaanxi Province of China under Grants 2022GY-057, 2021GY-086 and 2021GY-014, in part by the Key Industry Innovation Chain Projects of Shaanxi Province of China under Grants 2021ZDLGY07-04 and 2019ZDLGY13-01, and in part by a grant from the Youth Innovation Team of Shaanxi Universities.

References

[1] Q. Shi, J. Fan, Z. Wang, Z. Zhang, Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain, Pattern Recognit 130 (2022) 108837.
[2] M. Angelou, V. Solachidis, N. Vretos, P. Daras, Graph-based multimodal fusion with metric learning for multimodal classification, Pattern Recognit 95 (2019) 296–307.
[3] Y. Liu, L. Liu, Y. Guo, M.S. Lew, Learning visual and textual representations for multimodal matching and classification, Pattern Recognit 84 (2018) 51–67.
[4] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, L.-P. Morency, Multi-level multiple attentions for contextual multimodal sentiment analysis, in: 2017 IEEE International Conference on Data Mining, 2017, pp. 1033–1038.
[5] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, W. Kong, CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network,

in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021, pp. 5301–5311.
[6] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, 2019, p. 6558.
[7] D. Hazarika, R. Zimmermann, S. Poria, MISA: modality-invariant and -specific representations for multimodal sentiment analysis, in: MM '20: The 28th ACM International Conference on Multimedia, 2020, pp. 1122–1131.
[8] W. Yu, H. Xu, Z. Yuan, J. Wu, Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis, in: In Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 10790–10797.
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: 9th International Conference on Learning Representations, 2021.
[10] P. Malhotra, L. Vig, G. Shroff, P. Agarwal, et al., Long short term memory networks for anomaly detection in time series, in: 23rd European Symposium on Artificial Neural Networks, 2015, pp. 89–94.
[11] S. Bai, J.Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling (2018).
[12] R. Prabowo, M. Thelwall, Sentiment analysis: a combined approach, J Informetr 3 (2) (2009) 143–157.
[13] Y. Zhang, R. Jin, Z.-H. Zhou, Understanding bag-of-words model: a statistical framework, Int. J. Mach. Learn. Cybern. 1 (2010) 43–52.
[14] B. Li, T. Liu, Z. Zhao, P. Wang, X. Du, Neural bag-of-ngrams, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 3067–3074.
[15] P.-H. Chen, C.-J. Lin, B. Schölkopf, A tutorial on v-support vector machines, Appl Stoch Models Bus Ind 21 (2005) 111–136.
[16] I. Rish, et al., An empirical study of the naive bayes classifier, in: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001, pp. 41–46.
[17] S.J. Phillips, R.P. Anderson, R.E. Schapire, A brief tutorial on maxent, AT&T Research 190 (2005) 231–259.
[18] S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in: 2017 International Conference on Engineering and Technology, 2017, pp. 1–6.
[19] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.
[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: a robustly optimized bert pretraining approach, CoRR abs/1907.11692 (2019).
[21] W. Sheng, X. Li, Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network, Pattern Recognit 114 (2021) 107868.
[22] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th International Conference on Data Mining, 2016, pp. 439–448.
[23] A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, L.-P. Morency, Memory fusion network for multi-view sequential learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 5634–5641.
[24] O. Kampman, E.J. Barezi, D. Bertero, P. Fung, Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction, CoRR abs/1805.00705 (2018).
[25] A. Zadeh, M. Chen, S. Poria, E. Cambria, L. Morency, Tensor fusion network for multimodal sentiment analysis, in: EMNLP, Association for Computational Linguistics, 2017, pp. 1103–1114.
[26] S. Mai, H. Hu, S. Xing, Divide, conquer and combine: hierarchical feature fusion network with local and global perspectives for multimodal affective computing, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 481–492.
[27] P.P. Liang, Y.C. Lim, Y.H. Tsai, R. Salakhutdinov, L. Morency, Strong and simple baselines for multimodal utterance embeddings, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 2599–2609.
[28] Y.H. Tsai, P.P. Liang, A. Zadeh, L. Morency, R. Salakhutdinov, Learning factorized multimodal representations, in: 7th International Conference on Learning Representations, 2019.
[29] X. Yang, P. Molchanov, J. Kautz, Multilayer and multimodal fusion of deep neural networks for video classification, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 978–987.
[30] A. Agarwal, A. Yadav, D.K. Vishwakarma, Multimodal sentiment analysis via rnn variants, in: 2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering, 2019, pp. 19–23.
[31] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl Based Syst 161 (2018) 124–133.
[32] H. Pham, P.P. Liang, T. Manzini, L.-P. Morency, B. Póczos, Found in translation: Learning robust joint representations by cyclic translations between modalities, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 6892–6899.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 2017, pp. 5998–6008.
[34] W. Rahman, M.K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, E. Hoque, Integrating multimodal information in large pretrained transformers, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, 2020, p. 2359.
[35] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.-P. Morency, Words can shift: dynamically adjusting word representations using nonverbal behaviors, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 7216–7223.
[36] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 6–15.
[37] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[38] K. Ethayarajh, How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 55–65.
[39] M. Munezero, C.S. Montero, E. Sutinen, J. Pajunen, Are they different? affect, feeling, emotion, sentiment, and opinion detection in text, IEEE Trans Affect Comput 5 (2014) 101–111.
[40] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, 2019, pp. 13–23.
[41] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: pre-training of generic visual-linguistic representations, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020.
[42] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, COVAREP - A collaborative voice analysis repository for speech technologies, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4–9, 2014, IEEE, 2014, pp. 960–964.
[43] T. Drugman, M.R.P. Thomas, Detection of glottal closure instants from speech signals: a quantitative review, IEEE Trans. Speech Audio Process. 20 (3) (2012) 994–1006.
[44] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, CoRR abs/1606.06259 (2016).
[45] A. Zadeh, P. Pu, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), 2018.
[46] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A. Zadeh, L.-P. Morency, Efficient low-rank multimodal fusion with modality-specific factors (2018).
[47] Z. Sun, P. Sarma, W. Sethares, Y. Liang, Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 8992–8999.

Di Wang received the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2016. She is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. Her research interests include machine learning and multimedia information retrieval.

Xutong Guo received the B.S. degree in computer science and technology from Hebei University of Technology, Tianjin, China, in 2020. She is currently pursuing the M.S. degree with the computer science and technology at Xidian University. Her research interests focus on machine learning and multimodal sentiment analysis.

Yumin Tian received the B.Sc., and M.Sc. degrees in computer application from Xidian University, China, in 1984, and 1987, respectively. She is currently a Professor in the School of Computer Science and Technology, Xidian University. Her research interests include image processing, 3D shape recovery, digital watermarking, and computer vision.

Jinhui Liu received the B.S. degrees in information confrontation and Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 2006 and 2012. He is currently a Research Assistant with the school of computer science and technology in Xidian University. His research interest includes IC modeling and simulating technology, embedded system design and FPGA development.

Lihuo He received the B.Sc. degree in electronic and information engineering and PhD degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2008 and 2013. He is currently an associate professor with Xidian University. His research interests focus on computational vision, pattern recognition and artificial intelligence.

Xuemei Luo received the Ph.D. degree in computer system structure from Xidian University, Xi'an, China, in 2012. She is currently a lecturer with the School of Computer Science and Technology, Xidian University. Her research interests include color management, graphics and image processing, and machine learning.