Cross-Modal Enhancement Network for Multimodal Sentiment Analysis

Di Wang, *Member, IEEE*, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*—Multimodal sentiment analysis (MSA) plays an important role in many applications, such as intelligent question-answering, computer-assisted psychotherapy and video understanding, and has attracted considerable attention in recent years. It leverages multimodal signals including verbal language, facial gestures, and acoustic behaviors to identify sentiments in videos. Language modality typically outperforms nonverbal modalities in MSA. Therefore, strengthening the significance of language in MSA will be a vital way to promote recognition accuracy. Considering that the meaning of a sentence often varies in different nonverbal contexts, combining nonverbal information with text representations is conducive to understanding the exact emotion conveyed by an utterance. In this paper, we propose a Cross-modal Enhancement Network (CENet) model to enhance text representations by integrating visual and acoustic information into a language model. Specifically, it embeds a Cross-modal Enhancement (CE) module, which enhances each word representation according to long-range emotional cues implied in unaligned nonverbal data, into a transformer-based pre-trained language model. Moreover, a feature transformation strategy is introduced for acoustic and visual modalities to reduce the distribution differences between the initial representations of verbal and nonverbal modalities, thereby facilitating the fusion of distinct modalities. Extensive experiments on benchmark datasets demonstrate the significant gains of CENet over state-of-the-art methods.**

***Index Terms*—Multimodal sentiment analysis, pre-trained language model, transformer.**

I. Introduction

HUMANS naturally communicate with each other through multimodal signals such as language, audio and vision [1]. Therefore, understanding the emotion conveyed in an utterance requires a comprehensive understanding of different modalities. With the explosion of people-centric online videos, multimodal sentiment analysis (MSA) which leverages language (text), acoustic and visual modalities to identify human sentiments in videos has attracted much attention in recent years [1], [2]. Although many previous works have made great progress on the MSA task, achieving human-comparable performance remains challenging due to the large modality gap between heterogeneous modalities.

Transformer-based pre-trained language models [3]–[5] have achieved notable success in the natural language processing (NLP) field because of their strong contextual semantic feature extraction capacity and generality in downstream tasks through fine-tuning. Recent works have demonstrated that using pre-trained language models in MSA can greatly promote the recognition accuracy [2], [6]–[8]. Interaction Canonical Correlation Network (ICCN) [6] utilizes deep canonical correlation analysis to learn text-based audio and text-based video embeddings respectively and then fuses the two embeddings with text embedding. MISA [2] first utilizes three independent networks to learn three unimodal representations, then learns modality-invariant and modality-specific representations by multitask losses. Self-MM [7] proposes a self-supervised strategy to obtain unimodal labels, then learns inter-modal consistency and intra-modal specificity by a multitask framework based on the multimodal labels and the unimodal labels. The aforementioned works use BERT [3] as the text feature extractor and have achieved good results.

The pre-trained language models which are trained on a large textual corpus can greatly promote the understanding of sentiment in text modality. While for acoustic and visual modalities, feature extraction tools such as COVAREP [9] and Facet [1] are usually used to extract hand-crafted features first. And then a sequential neural network such as LSTM [10] is used to learn an utterance-level representation. Compared with text features learned by a pre-trained language model, nonverbal features are comparatively underdeveloped. Therefore, the language modality (verbal modality) usually performs better than the nonverbal modalities in MSA. For this reason, strengthening the

[1][Online]. Available: https://imotions.com/

Manuscript received 22 November 2021; revised 18 March 2022 and 18 May 2022; accepted 9 June 2022. Date of publication 16 June 2022; date of current version 30 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62072354, 61972302, 61876146, and 62072355, in part by the Key Research and Development Program of Shanxi Province under Grants 2022GY-057, 2019ZDLGY13-01, 2021GY-086, and 2021GY-014, in part by the Fundamental Research Funds for the Central Universities under Grants JB210305 and RW210419, and in part by the Youth Innovation Team of Shaanxi Universities . The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Ichiro Ide. *(Corresponding author: Quan Wang.)*

Di Wang, Quan Wang, and Yumin Tian are with the Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, the School of Computer Science and Technology, Xidian University, Xi'an 710071, China (e-mail: wangdi@xidian.edu.cn; qwang@xidian.edu.cn; ymtian@mail.xidian.edu.cn).

Shuai Liu is with the Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China (e-mail: sliu_4@stu.xidian.edu.cn).

Lihuo He and Xinbo Gao are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: lhhe@mail.xidian.edu.cn; xbgao@mail.xidian.edu.cn).

The implementation codes are available on https://github.com/Say2L/CENet.

Digital Object Identifier 10.1109/TMM.2022.3183830

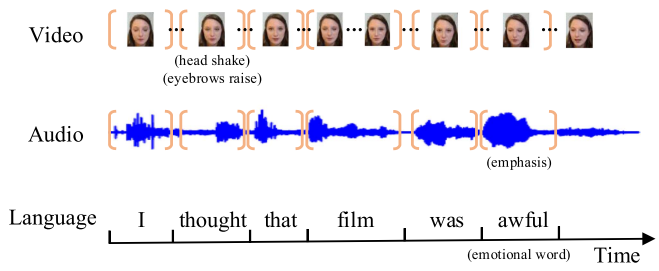

Fig. 1. An example of emotional information asynchrony among multimodalities.

importance of language modality in MSA will be a feasible way to improve the recognition accuracy of MSA. In addition, considering that a sentence can have different meanings in diverse nonverbal contexts, such as facial expressions and tones of voice, integrating nonverbal information into a language model can help to understand the emotion conveyed in an utterance exactly.

In this paper, we propose a Cross-modal Enhancement Network (CENet) model to enhance text representations by incorporating emotional information of visual and acoustic modalities. The core of CENet is the Cross-modal Enhancement (CE) module which can be embedded in a transformer-based pre-trained language model. The CE module consists of two parts: a cross-modal embedding unit and an enhancement embedding gate. The cross-modal embedding unit focuses on capturing long-range nonverbal emotional cues from unaligned nonverbal data. The long-range nonverbal emotional cues refer to emotional information extracted from nonverbal frames with different timestamps. Generally, relevant information in language, visual and acoustic modalities is asynchronous [11]. As shown in Fig. 1, "awful" is an emotional word, and the informative facial expression appears at a different moment from saying the word. Cross-modal attention [11], which is based on self-attention, can capture long-range dependencies from unaligned multimodal data. Therefore, we use it to obtain asynchronous nonverbal emotional contexts for each word from audio and vision respectively. We call these visual and acoustic emotional contexts text-based visual and acoustic embeddings, respectively. The enhancement embedding gate utilizes the text-based nonverbal embeddings to generate a nonverbal enhancement embedding that contains long-range nonverbal emotional contexts. Then the enhancement embedding will update the text representation within a pre-trained language model.

Considering that there are large distribution differences between initial verbal and nonverbal representations, we introduce a feature transformation strategy to reduce the distribution gap. Inspired by the token index sequence derived from textual vocabulary, we cluster the hand-crafted features of acoustic and visual frames to establish "acoustic vocabulary" and "visual vocabulary," respectively. Then, we can obtain the index sequences of acoustic and visual modalities by querying the nonverbal vocabularies. In this way, the high dimensional nonverbal feature is transformed into an individual index which is similar to a word token. Therefore, the initial distribution gaps between heterogeneous modalities are diminished, which will further narrow the distribution differences between verbal and nonverbal features at the stage of fusing. Thus, the feature transformation strategy can facilitate the integration of text representations and nonverbal emotional contexts.

The contributions of this work can be summarized as follows:

- A Cross-modal Enhancement Network is proposed to enhance the text representation within a pre-trained language model by incorporating long-range nonverbal emotional contexts.
- A feature transformation strategy is proposed to facilitate the integration of different modalities by reducing the distribution differences between the initial representations of verbal and nonverbal modalities.
- Extensive experiments on two benchmark multimodal sentiment analysis datasets demonstrate that the proposed method greatly outperforms state-of-the-art methods.

II. Related Work

In this section, we first introduce some closely related works on unimodal sentiment analysis and multimodal sentiment analysis. Then the pre-trained language models will be discussed.

A. Unimodal Sentiment Analysis

1) Textual Sentiment Analysis: Text-based sentiment analysis is an active and successful research area [12]–[14]. Early works usually present a text utterance using bag-of-words, and then utilize a machine learning method such as SVM to classify sentiment polarity [15], [16]. With the development of deep learning, CNN, Recursive Neural Networks (RNN) and Long-Short Term Memory (LSTM) are widely used in textual sentiment analysis [17]–[20]. Recently, transformer [21], which is based on self-attention, has emerged in many areas of NLP. Compared with recurrent and convolutional layers, the self-attention layer is more computationally efficient and easier to learn long-range dependencies [21]. Transformer-based pre-trained language models designed for sentiment analysis [22]–[25] have greatly facilitated the development of textual sentiment analysis.

2) Visual Sentiment Analysis: Visual sentiment analysis can be divided into image and video sentiment analysis. For image sentiment analysis, fine-tuning pre-trained Convolutional Neural Networks (CNN) is adopted by many works to obtain emotion-related features [26]–[30]. For video sentiment analysis, visual information is distributed in time and space. Generally, 3D Convolutional Neural Networks (C3D) or CNN followed by LSTM are applied to extract emotion-related features from spatial-temporal visual input [31]–[33].

3) Acoustic Sentiment Analysis: For acoustic sentiment analysis, the functionals of acoustic low-level descriptors (LLD), which are hand-crafted and built upon prior knowledge, are used as acoustic features for emotion recognition in previous works [34], [35]. Recent studies learn acoustic features directly from raw audio signals using Deep Neural Network (DNN) [36]–[39]. These acoustic features are then fed to a sequential neural network such as LSTM to capture temporal dynamics.

B. Multimodal Sentiment Analysis

Previous works in MSA can be divided into two categories in light of whether pre-trained language models are used. The first kind of method does not utilize pre-trained models. These methods often use GloVe [40] word embeddings followed by LSTM [10] to extract language representations. The early work [41] concatenates text, visual and acoustic representations obtained by different feature extraction networks, then utilizes multiple kernel learning (MKL) as the classifier to predict sentiment intensity. Convolutional recurrent multiple kernel learning (CRMKL) [42] is an improvement version of [41]. It combines CNN and RNN to extract spatial and temporal features from visual data and also applies MKL as the classifier. However, directly combining distinct unimodal representations may cause information loss. To avoid this problem, subsequent works often focus on designing elaborate fusion frameworks. Tensor fusion network (TFN) [43] learns intra-modality dynamics through a modality embedding subnetwork and obtains inter-modality interactions by calculating outer-product. Despite TFN can fuse multimodal information well, it has an expensive computational cost. Low-rank multimodal fusion (LMF) [44] reduces the computational cost of TFN by low-rank tensor. Recurrent attended variation embedding network (RAVEN) [45] utilizes fine-grained non-verbal subword information to dynamically adjust word representations for multimodal fusion. Factorized multimodal transformer (FMT) [46] applies factorized multimodal self-attention (FMS) to build inter-modal interactions. The FMS takes into account all factors in the combination of three modalities as input. Multimodal transformer (MulT) [11] applies cross-modal attention to translate one modality to another modality and vice versa, thus building interactions between different modalities. Different from MulT, our method utilizes long-range emotional cues captured by cross-modal attention to enhance text representations. In other words, the information flow is unidirectional, from nonverbal modalities to text modality.

The second kind of method utilizes pre-trained language models to extract text features and usually obtains better results than the first kind of method. Interaction canonical correlation network (ICCN) [6] builds interactions between text and nonverbal modalities by deep canonical correlation analysis. MISA [2] learns an invariant representation and a specific representation for each modality through four distinct loss functions, and then fuses different representations to predict sentiment. Self-supervised multi-task multimodal sentiment analysis network (Self-MM) [7] introduces a self-supervised label generation module to obtain extra unimodal labels. Then it acquires information-rich unimodal representations for MSA by joint learning one multimodal task and three unimodal subtasks. CM-BERT [47] utilizes masked multimodal attention to dynamically adjust the weight of each word-level feature of the output of BERT. And it only exploits the information of acoustic modality to adjust text features. Unlike CM-BERT, our CENet adds the nonverbal enhancement embedding obtained by the CE module into the text representation of the output of a middle layer of a pre-trained language model. And CENet enhances text representations by leveraging the emotional information of both acoustic and visual modalities. MAG-BERT [8] introduces a multimodal adaptation gate that enables BERT [3] to accept the representations of nonverbal modalities. In this paper, we also integrate acoustic and visual information into a transformer-based language model. However, our method can capture asynchronous emotion cues from unaligned nonverbal data to enhance text representations, MAG-BERT can only process word-level aligned multimodal data.

C. Pre-Trained Language Models

Pre-trained language models have superior performance compared with GloVe [40] in word representation. ELMo [48] pre-trains a bi-directional LSTM on a large-scale unsupervised language corpus. GPT [49] is a transformer-based model that performs better in capturing long-range dependencies compared to ELMo. However, they are both unidirectional language models which limit the further improvement of contextual representation capability.

BERT [3] is a bi-directional pre-trained language model based on transformer [21]. By using a masked language model (MLM) pre-training objective, BERT acquires bidirectional context perception ability. XLNet is also a transformer-based bi-directional pre-trained language model, which avoids the inconsistency in the pre-training and fine-tuning stages of MLM by building a permutation language model in the pre-training stage. Most of the existing pre-trained language models are variants of BERT and XLNet. VideoBERT [50] is a variant of BERT, which applies vector quantization to make the BERT model jointly accept sequences of visual and linguistic tokens. And it learns bidirectional joint distributions of video and text by text-only, video-only and text-video pre-training objectives. SentiLARE [25] is a transformer-based pre-trained model designed specifically for the sentiment analysis task. It inherits from RoBERTa [5] and integrates linguistic knowledge including part-of-speech tag and sentiment polarity into the language model.

Similar to VideoBERT, we apply vector quantization in our feature transformation strategy for nonverbal data. The motivation of our feature transformation strategy is to diminish the distributional differences between heterogeneous modalities. While VideoBERT uses the vector quantization to encourage the model to focus on high-level semantics and longer-range temporal dynamics in video rather than the low-level properties such as local textures and motions. Additionally, VideoBERT combines word tokens with visual tokens and utilizes BERT to directly learn multimodal representation. While our CENet first exploits an extra module to extract nonverbal features and then integrates the nonverbal features into a pre-trained language model. In this paper, We use SentiLARE, BERT and XLNet as the language model respectively to comprehensively assess the CENet framework.

III. Cross-Modal Enhancement Network

In this section, we first present the MSA task setup. Then, we introduce the feature transformation strategy and the cross-modal enhancement module. Finally, the overall architecture of the CENet model is described.

Algorithm 1: Feature Transformation Strategy.

Input: nonverbal frame set of F_m with the size of $N_m \times d_m$, nonverbal query feature sequence S_m with the size of $L_m \times d_m$, the number of clusters K_m, the maximum number of iteration M;
Output: nonverbal index sequence I_m;

```
Initialize cluster centers C_m with the size of K_m × d_m, randomly;
Initialize cluster indexes of all training frames T_m ← −1;
repeat
  for i = 0, 1, ..., N_m − 1 do
    T_m[i] ← updateClusterIndex(F_m[i], C_m);
  end for
  for i = 0, 1, ..., K_m − 1 do
    C_m[i] ← updateClusterCenter(T_m, F_m);
  end for
until T_m no longer changes or reaching the maximum number of iteration.
for i = 0, 1, ..., L_m − 1 do
  I_m[i] ← queryClusterIndex(S_m[i], C_m);
end for
return I_m
```

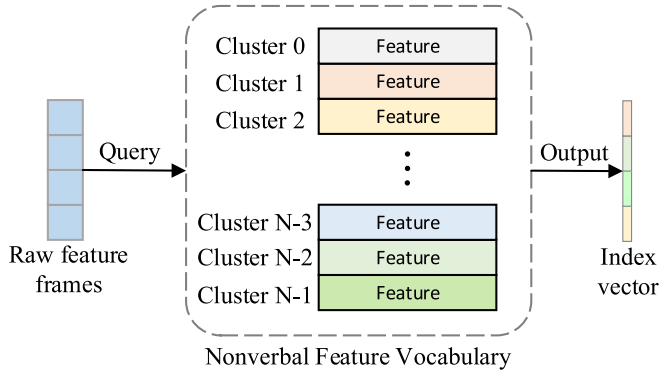

Fig. 2. Feature transformation strategy. The nonverbal feature vocabulary is built by clustering algorithm on training dataset. By querying the vocabulary, a raw feature frames segment can be transformed to a cluster index sequence.

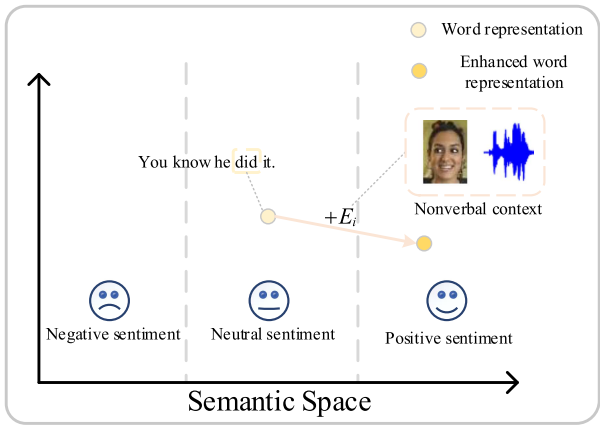

Fig. 3. An illustration of updating word representation by the cross-modal enhancement embedding. E_i denotes the i-th vector in nonverbal enhancement embedding E.

A. Task Setup

Multimodal sentiment analysis is to detect sentiment in a video segment by using multimodal signals. For a video segment X, it is composed of three parts, including text (t), acoustic (a) and visual (v) sequences, which are represented as $X_m = \{x_m^1, x_m^2, \ldots, x_m^{L_m}\}$, $m \in \{t, a, v\}$, respectively. Here, L_m represents the sequence length of modality m. x_t^i, x_a^i and x_v^i denote the i-th word, audio frame and visual frame, respectively. For text sequence X_t, we use the tokenizer of a pre-trained language model to obtain its corresponding token sequence $I_t \in \mathbb{R}^{L_t}$. For acoustic and visual sequences, we extract their raw features by Facet and COVAREP [9], respectively. The raw acoustic and visual feature sequences are represented as $S_a \in \mathbb{R}^{L_a \times d_a}$ and $S_v \in \mathbb{R}^{L_v \times d_v}$, where d_a and d_v are acoustic and visual feature dimensions, respectively. The details of extracting nonverbal features are discussed in Section Experimental Settings.

Given a video segment X, the goal of CENet is to predict the sentiment intensity y of X, where $y \in [-3, 3]$ is a continuous intensity variable with $y > 0$ denotes positive sentiment, $y < 0$ denotes negative sentiment and $y = 0$ denotes neutral sentiment.

B. Feature Transformation for Nonverbal Modalities

For pre-trained language models, the initial textual representation is a word index sequence from a vocabulary. However, the representations of visual and acoustic are real vector sequences. Therefore, we propose a feature transformation strategy that transforms nonverbal vectors into indices to diminish the initial distribution differences between heterogeneous modalities, which will further narrow the distribution gaps of verbal and nonverbal features when fusion. Thus, it will facilitate the integration of text representations and nonverbal emotional contexts.

The feature transformation strategy utilizes an unsupervised clustering algorithm to establish "acoustic vocabulary" and "visual vocabulary," respectively. By querying the nonverbal vocabulary, a raw feature sequence can be transformed into an index sequence. Fig. 2 shows an illustration of the feature transformation process.

Due to the merits of low computational complexity and simplicity of k-means method, we use k-means to learn vocabularies of nonverbal modalities. Without loss of generality, other clustering and dictionary learning methods can also be used for learning acoustic and visual vocabularies.

Specifically, we first gather all the frames of acoustic and visual segments from training set to form two frame sets $F_m \in \mathbb{R}^{N_m \times d_m}, m \in \{a, v\}$, N_m denotes the number of frames for modality m. Then, the frames in F_m are divided into k_m groups by k-means as follows:

$$C_m = \text{Kmeans}(F_m), \tag{1}$$

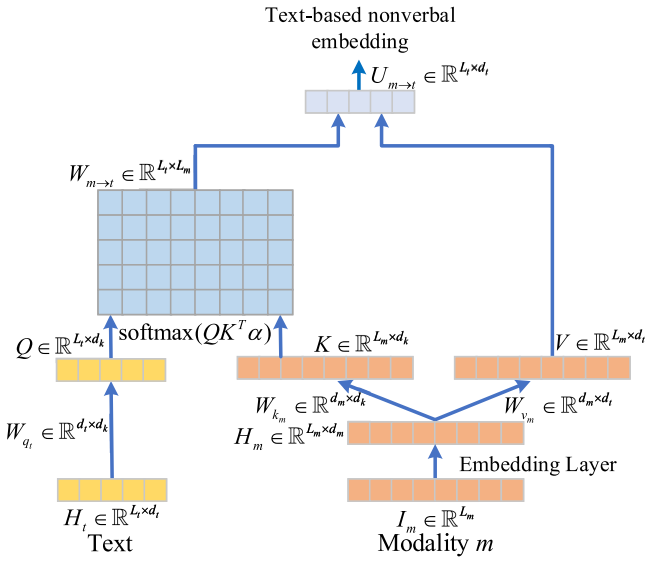

Fig. 4. An illustration of cross-modal embedding unit.

where $C_m = \{c_m^0, c_m^1, \ldots, c_m^{K_m-1}\}$, c_m^j is the j-th cluster center of modality m. Then we can build "visual vocabulary" and "acoustic vocabulary" based on the cluster centers C_a and C_v, respectively. Given a nonverbal feature sequence $S_m = \{s_m^0, s_m^1, \ldots, s_m^{L_m-1}\} \in \mathbb{R}^{L_m \times d_m}$, L_m is the length of the sequence and s_m^i is the feature of the i-th frame of the sequence. By querying the nonverbal vocalubary, we can get a corresponding index sequence $I_m \in \mathbb{R}^{L_m}$. For the feature of i-th frame s_m^i in S_m, the querying process is calculated as follows:

$$l_m^i = \underset{\mathrm{j}}{\operatorname{argmin}}(||s_m^i - c_m^j||_2), j = 0, 1, \ldots, K_m - 1, \quad (2)$$

where l_m^i is the index label of the i-th frame s_m^i. Subsequently, the obtained index sequence $I_m = \{l_m^0, l_m^1, \ldots, l_m^{L_m-1}\}$ will serve as the representation of modality m. The process of the feature transformation strategy is shown in Algorithm 1.

Note that the raw nonverbal features are mainly content irreverent emotional features. Therefore, the trivial solution that dividing frames with similar content but different emotions into one cluster will be avoided.

C. Cross-Modal Enhancement Module

The proposed CE module integrates long-range visual and acoustic information into a pre-trained language model to enhance text representations. The architecture of the CE module is shown in Fig. 5.

The key component of the CE module is the cross-modal embedding unit which is illustrated in Fig. 4. It utilizes cross-modal attention to capture long-range nonverbal emotional information and generates text-based nonverbal embeddings.

Specifically, given an index vector $I_m \in \mathbb{R}^{L_m}$, $m \in \{a, v\}$, the cross-modal embedding unit first inputs it to an embedding layer:

$$H_m = \text{Embedding}(I_m) \in \mathbb{R}^{L_m \times d_m}, \quad (3)$$

where H_m is the output of the embedding layer, and d_m denotes the embedding dimension. The parameters of the embedding layer are learnable. The role of the embedding layer is to map the nonverbal index vector, which is obtained by the feature transformation strategy, to a high-dimensional space.

Given weight matrices $W_{Q_t} \in \mathbb{R}^{d_t \times d_k}$, $W_{K_m} \in \mathbb{R}^{d_m \times d_k}$, $W_{V_m} \in \mathbb{R}^{d_m \times d_t}$, $m \in \{a, v\}$, the queries Q, keys K and values V in cross-modal attention are calculated as

$$Q = H_t W_{Q_t} \in \mathbb{R}^{L_t \times d_k}, \quad (4)$$

$$K = H_m W_{K_m} \in \mathbb{R}^{L_m \times d_k}, \quad (5)$$

$$V = H_m W_{V_m} \in \mathbb{R}^{L_m \times d_t}, \quad (6)$$

where $H_t \in \mathbb{R}^{L_t \times d_t}$ is the text hidden representation which is the output of a middle layer of a pre-trained language model, and d_t denotes the text feature dimension. The definitions of Q, K and V in cross-modal attention are similar to those in self-attention [21].

Afterward, we can get the attention weight matrix of text modality to nonverbal modalities by

$$\begin{aligned} W_{m\to t} &= \text{softmax}\left(QK^{\mathsf{T}}\alpha\right) \\ &= \text{softmax}\left(H_t W_{Q_t} W_{K_m}^{\mathsf{T}} H_m^{\mathsf{T}} \alpha\right), \end{aligned} \quad (7)$$

where $W_{m\to t} \in \mathbb{R}^{L_t \times L_m}, m \in \{a, v\}$ is an attention weight matrix. The (i, j)-th element of $W_{m\to T}$ indicates the attention of the i-th word of text modality to the j-th frame of modality m. α is a scaling parameter. At the initial training stage, as verbal representations and nonverbal representations lie in two different feature spaces, the correlations between verbal and nonverbal representations will be small. In this way, elements in the weight matrix will be also small. For better learning model parameters, we use a hyper-parameter α to scale the matrix before softmax processing.

Based on the attention weight matrix $W_{m\to t}$, we can obtain the text-based nonverbal embedding $U_{m\to t}$ as follows:

$$U_{m\to t} = W_{m\to t} V \in \mathbb{R}^{L_t \times d_t}. \quad (8)$$

The obtained text-based nonverbal embedding can be regarded as the latent emotional information in modality m chosen by text modality.

Combining the text-based acoustic embedding $U_{a\to t}$ and the text-based visual embedding $U_{v\to t}$, we can obtain a nonverbal enhancement embedding by

$$E = \text{Gate}(U_{a\to t}; U_{v\to t}), \quad (9)$$

where ";" denotes the concatenate operation, and $\text{Gate}(\cdot)$ is the enhancement embedding gate which is composed of fully connected dense layers. The role of the enhancement embedding gate is to fuse the text-based acoustic embedding $U_{a\to t}$ with the text-based visual embedding $U_{v\to t}$ and generate the nonverbal enhancement embedding E.

Finally, the text representation H_t will be updated by the nonverbal enhancement embedding E as

$$H_t' = H_t + E. \quad (10)$$

In general, every word in an utterance has a nonverbal context. The meaning of a same word can be varied in different nonverbal

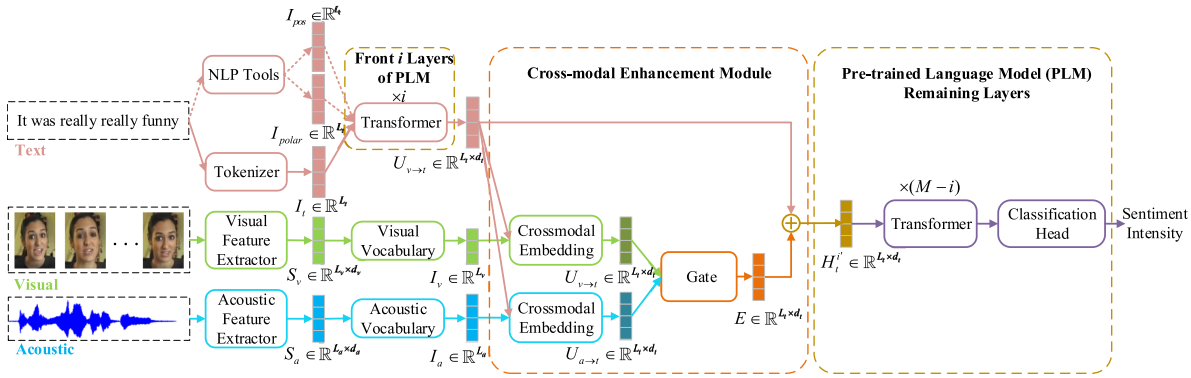

Fig. 5. Overall architecture of CENet. A cross-modal enhancement module is embedded between the i-th and $(i+1)$-th layers of a transformer-based pre-trained language model (SentiLARE is used here).

contexts. Furthermore, the nonverbal context is not strictly synchronized with the word, seeing an example in Fig. 1. Therefore, we propose the CE module to generate the nonverbal enhancement embedding E which provides nonverbal context information for text. By adding the nonverbal enhancement embedding E to the text representation, we can adjust the text representation to be more semantically accurate. As shown in Fig. 3, "did" in the sentence "You know he did it" is a neutral word, while it can be changed to a positive word in the context of a smiling expression and an excited voice.

D. Overall Architecture of CENet

The main architecture of CENet is incorporating the proposed cross-modal enhancement module into a transformer-based pre-trained language model. Fig. 5 shows the overall architecture of CENet in which we use the SentiLARE network [25] as the language model. SentiLARE utilizes word linguistic knowledge including part-of-speech and word emotional polarity to learn sentiment-aware language representations. Supposing that the CE module is integrated into the i-th layer of a pre-trained language model. Algorithm 2 shows the CENet during the training stage. It is worth noting that any transformer-based pre-trained language model can be integrated with our CE module.

Following the settings of SentiLARE, given a word sequence X_t, we first learn its part of speech sequence I_{pos} by the Stanford Log-Linear Part-of-Speech (POS) Tagger [51] and a word-level sentiment polarity sequence I_{polar} by SentiwordNet [52]. Then we obtain its token index sequence I_t by the tokenizer of the pre-trained language model. With the input of sequences I_t, I_{pos} and I_{polar}, the initial text representation with language knowledge enhancement can be obtained as

$$H_t^0 = \text{BertEmbedding}(I_t) + \text{Embedding}(I_{pos}) + \text{Embedding}(I_{polar}), \tag{11}$$

where BertEmbedding denotes the embedding operations in BERT [3] and Embedding denotes a embedding layer. Then, H_t^0 will go through $1 \rightarrow i$ transformer layers. We refer to H_t^i as the output of the i-th layer. The CE module is embedded between the i-th and $(i+1)$-th layers of a pre-trained language model. The inputs of the CE module are the nonverbal index vectors I_a and I_v and the text hidden representation H_t^i. It produce a enhancement embedding E. Afterward, the text representation H_t^i will be updated by the enhancement embedding E as

$$H_t^{i'} = H_t^i + E. \tag{12}$$

Next, the updated text representation $H_t^{i'}$ will be the input of the $(i+1)$-th layer and pass through the remaining $M-i$ layers, where $M=12$ in SentiLARE. For the M-th layer, its output h_{cls} will be a text-dominated high-level sentiment representation with visual and acoustic information.

At the final step, the text representation h_{cls} is input into a classification head to obtain the sentiment intensity $\hat{y}$.

IV. Experiments

A. Datasets

We evaluate the CENet model on two public benchmark multimodal sentiment analysis datasets: CMU-MOSI [14] and CMU-MOSEI [1].

1) CMU-MOSI: It is a multimodal dataset including text, visual and acoustic modalities. It is derived from 93 Youtube movie review videos. These videos are clipped into 2,199 segments. Each segment is annotated with a sentiment intensity in the range of [-3, 3]. The dataset is divided into three parts, training set (1,284 segments), validation set (229 segments) and test set (686 segments).

2) CMU-MOSEI: It is similar to CMU-MOSI but on a larger scale. It contains 23,453 annotated video segments from online video-sharing websites covering 250 different topics and 1000 distinct speakers. Samples in CMU-MOSEI are labeled with both sentiment intensity in the range of [-3, 3] and 6 basic emotions. Hence, CMU-MOSEI can be used for sentiment analysis and emotion recognition tasks.

In the following sections, let MOSI and MOSEI represent CMU-MOSI and CMU-MOSEI, respectively.

B. Feature Extraction

1) Language Features: Most of the existing works use word embeddings either from GloVe or from a pre-trained language model. In this paper, we use a pre-trained model to obtain word

Algorithm 2: Crossmodal Enhancement Network.

Input: nonverbal index sequences I_a and I_v, text token index sequence I_t, multimodal sentiment label y;
Output: prediction sentiment intensity $\hat{y}$;
Initialize the pre-trained language model parameters $M(\theta; x)$;
Initialize visual and acoustic embedding layer parameters $Embedding_v(\theta_v; x)$ and $Embedding_a(\theta_a; x)$, respectively;
Initialize cross-modal attention parameters W_{Q_a}, W_{K_a}, W_{V_a} and W_{Q_v}, W_{K_v}, W_{V_v}.
Initialize enhancement embedding gate parameters $Gate(\theta_{gate}; x)$;
for $epoch = 1, 2, \ldots, end$ **do**
  for mini-batch in dataLoader **do**
    Compute the text representation H_t^i which is the output of the i-th layer of a pre-trained language model;
    for $m = a, v$ **do**
      $H_m = Embedding_m(I_m)$;
    end for
    Compute the text-based visual embedding $U_{v \to t}$ and the text-based acoustic embedding $U_{a \to t}$ using Equation (4∼8);
    Compute the nonverbal enhancement embedding E using Equation (9);
    Compute the enhanced text representation $H_t^{i'}$ using Equation (10);
    Input $H_t^{i'}$ to the remaining pre-trained language model layers;
    Compute the prediction sentiment intensity $\hat{y}$;
    Compute loss using L2 loss function;
    Update the network parameters using BP algorithm;
  end for
end for

embeddings. In addition to the common BERT and XLNet, we also use SentiLARE as the language model. Compared with BERT and XLNet, SentiLARE adds a part-of-speech (POS) embedding and a word-level sentiment polarity embedding for each sentence. The POS and word-level sentiment polarity are derived from Stanford Log-Linear Part-of-Speech Tagger [51] and SentiwordNet [52] respectively.

2) Visual Features: For the MOSI dataset, we use the Py-Feat toolkit [53] to extract the facial features of each visual frame including facial landmarks, 7 facial expressions and facial action units. The facial landmarks are used to select frames with opened eyes by calculating the aspect ratio of the eyes. Frames with closed eyes may cause uncertainty in expression analysis and are therefore discarded. Finally, a 27-dimensional feature which contains facial expressions and facial action units is formed for each frame with opened eyes. For the MOSEI dataset, Facet is used to extract visual features with 35-dimensions containing facial action units and face pose.

3) Acoustic Features: COVAREP software [9] is used to extract acoustic features, which are related to emotions and tones of speech, including 12 Mel-frequency cepstral coefficients, pitch, voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. The feature dimension is 74 for both MOSI and MOSEI datasets.

C. Baselines

The proposed CENet is compared with the following baselines:

1) Unimodal: For acoustic and visual modalities, a two-layer Transformer is used to extract emotional representation from hand-crafted features. For text modality, we set three baselines including two-layer Transformer using GloVe word embeddings, BERT, and SentiLARE.

2) TFN: Tensor Fusion Network (TFN) [43] learns intra-modality and inter-modality dynamics through modality embedding subnetwork and a novel fusion method called tensor fusion, respectively.

3) LMF: Low-Rank Multimodal Fusion (LMF) [44] applies a fusion strategy similar to TFN but uses a low-rank tensor to reduce computational cost.

4) RAVEN: Recurrent Attended Variation Embedding Network (RAVEN) [45] considers fine-grained nonverbal subword sequences and dynamically adjusts word features using corresponding nonverbal features.

5) FMT: Factorized Multimodal Transformer (FMT) [46] introduces a new transformer architecture for multimodal sequential learning which is based on Factorized Multimodal Self-attention (FMS). FMS takes into account all the factors of the combination of the three modalities.

6) MulT: Multimodal Transformer (MulT) [11] utilizes directional pairwise cross-modal attentions to build interactions between modalities. It does so by translating one modality to a target modality and vice versa.

7) CM-BERT: CM-BERT [47] exploits acoustic emotional information to dynamically adjust the weight of each word-level feature of the output of BERT.

8) MISA: Modality-Invariant and -Specific Representations (MISA) [2] learns an invariant representation and a specific representation for each modality through a multi-task learning model.

9) Self-MM: Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning (Self-MM) [7] designs a self-supervised label generation module to obtain extra unimodal labels. Then it combines unimodal and multimodal tasks to learn inter-modal consistency and intra-modal specificity.

10) MAG-BERT: The Multimodal Adaptation Gate for BERT (MAG-BERT) [8] incorporates aligned nonverbal information into the text representation within BERT pre-trained model.

For a fair comparison, we retrain some baselines using the same nonverbal features as ours. Furthermore, we compare our CENet(S) with SOTA methods including MISA(S) and

TABLE I
SENTIMENT PREDICTION RESULTS ON MOSI AND MOSEI. FOR ACC-2 AND F1, THE VALUES ON THE LEFT AND RIGHT SIDES OF "/" ARE MEASURES CALCULATED BASED ON [54] AND [11], RESPECTIVELY. "-" INDICATES THE RELATED VALUE IS NOT GIVEN IN THE CORRESPONDING REFERENCE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. (B) AND (S) INDICATE THAT THE LANGUAGE FEATURES ARE EXTRACTED BY BERT AND SENTILARE, RESPECTIVELY. (G) MEANS THE GLOVE WORD EMBEDDINGS ARE USED. [1, 2, 3, AND 4] REPRESENT THE CORRESPONDING RESULTS ARE DERIVED FROM REFERENCES [2], [46], [47], AND [8], RESPECTIVELY

Modality	Model	MOSI				MOSEI				Data State
		MAE	Corr	Acc-2	F1	MAE	Corr	Acc-2	F1	
Unimodal	A	1.403	0.112	55.83/57.01	54.51/55.88	0.820	0.236	65.55/63.90	64.94/61.98	N/A
	V	1.372	0.160	57.14/58.23	56.39/57.65	0.823	0.211	70.53/63.48	61.31/52.14	N/A
	T(G)	1.188	0.455	67.93/70.12	66.59/69.03	0.748	0.507	76.62/74.76	75.59/73.18	N/A
	T(B)	0.722	0.789	83.41/85.62	83.70/85.76	0.560	0.757	82.25/85.50	82.66/85.50	N/A
	T(S)	0.580	0.864	87.32/89.94	87.19/89.88	0.537	0.793	85.05/87.23	84.70/87.24	N/A
Multimodal	TFN(G)[1]	0.970	0.633	73.90/-	73.40/-	-	-	-	-	Unaligned
	LMF(G)[1]	0.912	0.668	76.40/-	75.70/-	-	-	-	-	Unaligned
	RAVEN(G)[1]	0.915	0.691	78.00/-	76.60/-	0.614	0.662	79.10/-	79.50/-	Aligned
	MulT(G)[1]	0.871	0.698	-/83.00	-/82.80	0.580	0.703	-/82.50	-/82.30	Aligned
	FMT(G)[2]	0.837	0.744	81.50/83.50	81.40/83.50	-	-	-	-	Aligned
	CM-BERT(B)[3]	0.729	0.791	-/84.50	-/84.50	-	-	-	-	Aligned
	MISA(B)	0.721	0.788	83.63/85.83	83.85/85.97	0.547	0.772	82.12/85.79	81.75/85.90	Unaligned
	MAG-Bert(B)[4]	0.712	0.796	84.20/86.10	84.10/86.00	-	-	-/84.7	-/84.5	Aligned
	Self-MM(B)	0.704	0.803	83.93/86.00	84.05/86.06	0.530	0.770	82.40/85.07	82.02/85.12	Unaligned
	MISA(S)	0.605	0.848	86.09/88.52	86.22/88.58	0.524	0.795	84.74/86.76	84.57/86.90	Unaligned
	Self-MM(S)	0.590	0.864	86.88/89.48	87.01/89.54	0.517	0.794	84.79/86.25	84.58/86.32	Unaligned
	MAG-SentiLARE(S)	0.573	0.867	87.45/89.62	87.35/89.57	0.528	**0.796**	84.65/87.09	84.98/87.08	Aligned
	CENet(S)	**0.570**	**0.870**	**88.63/90.85**	**88.55/90.82**	**0.515**	**0.796**	**85.49/87.86**	**85.21/87.82**	Unaligned

Self-MM(S), MAG-SentiLARE(S) which also use SentiLARE as their language model.

D. Evaluation Metrics

Following prior works [2], [7], [8], we build two evaluation tasks including binary classification and regression. For binary classification, the binary classification accuracy (Acc-2) and the weighted F1 score (F1) are reported. There are two classification ways: negative/non-negative classification and negative/positive classification. For regression, the recognition performance is evaluated by the mean absolute error (MAE) and the Pearson correlation (Corr).

E. Parameter Settings

The proposed CENet is trained by the Adam optimizer with learning rates between $\{1e-4, 2e-5, 4e-5, 6e-5\}$. The cluster number is set to 16 and the scaling parameter α in the CE module is set to 8. The CE module is embedded between the first and the second layers of a pre-trained language model unless otherwise specified. The number of fully connected dense layers in the enhancement embedding gate is one. All models use the validation set of MOSI to find the most appropriate hyper-parameters.

F. Comparison With Baselines

The sentiment analysis results of CENet(S) and baselines on MOSI and MOSEI are presented in Table I. CENet(S) model utilizes the SentiLARE pre-trained model which is specially designed for emotional language processing tasks. For a fair comparison, we retrain the baseline models including MISA(S), Self-MM(S) and MAG-SentiLARE(S) which use SentiLARE as their language model. In the experiments, only methods designed for word-level aligned datasets are performed on aligned datasets, other comparison methods and CENet(S) are conducted on unaligned datasets.

From Table 1, we have the following observations.

- For unimodal baselines, text modality outperforms non-verbal modalities. SentiLARE gives the best results on text modality, followed by BERT. We argue that there may be several reasons for this. First, most people prefer language to express their emotions. Second, in the field of multimodal sentiment analysis, the methods on nonverbal modalities are underdeveloped compared with pre-trained language models on text modality.
- The results of MISA(S), Self-MM(S) and MAG-SentiLARE(S) are better than those of MISA(B), Self-MM(B) and MAG-BERT(B), respectively. This indicates that the introduction of SentiLARE can further improve the accuracy of SOTA methods. However, compared with the results of SentiLARE on text modality, these multimodal methods do not perform better. Possible reasons will be discussed in a later section.
- CENet(S) outperforms baselines significantly on both MOSI and MOSEI. And the multimodal methods including MISA(S), Self-MM(S) and MAG-SentiLARE(S) use SentiLARE as their language model, CENet(S) still outperforms them greatly. Compared with SentiLARE, CENet(S) achieves large performance improvements, which reflects the effectiveness of enhancing the text representation within a pre-trained language model using long-range non-verbal dependencies.

The MAE and Corr of human performance results on MOSI are 0.61 and 0.83 respectively [14]. The performance of CENet on MOSI even exceeds that of humans.

TABLE II
SENTIMENT PREDICTION RESULTS ON MOSI. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. (B) AND (X) INDICATES THAT THE LANGUAGE FEATURES ARE EXTRACTED BY BERT AND XLNET, RESPECTIVELY. T REPRESENTS THAT ONLY TEXT MODALITY IS USED

Model	MOSI			
	Acc-2	F1	MAE	Corr
T(B)	83.41/85.62	83.70/85.76	0.722	0.789
TFN(B)	82.66/84.31	82.71/84.31	0.738	0.765
LMF(B)	82.81/84.64	82.84/84.61	0.742	0.783
MulT(B)	81.55/83.96	81.81/84.12	0.735	0.777
MISA(B)	83.63/85.83	83.85/85.97	0.721	0.788
Self-MM(B)	83.93/86.00	84.05/86.06	0.704	0.803
MAG-BERT(B)	84.23/86.11	84.17/86.10	0.720	0.792
CENet(B)	**84.40/86.74**	**84.24/86.66**	**0.698**	**0.806**
T(X)	84.84/87.04	84.76/87.04	0.671	0.821
TFN(X)	84.22/86.27	84.30/86.30	0.699	0.801
LMF(X)	83.59/85.62	83.74/85.70	0.696	0.805
MulT(X)	84.08/86.14	84.20/86.20	0.721	0.804
MISA(X)	83.93/86.31	84.10/86.40	0.712	0.803
Self-MM(X)	84.23/86.92	84.33/86.94	0.671	0.817
MAG-XLNet(X)	85.11/87.33	85.01/87.29	0.678	0.819
CENet(X)	**86.15/88.41**	**86.06/88.38**	**0.648**	**0.832**

G. CENet With Other Pre-Trained Language Models

For comprehensively demonstrating the effectiveness and generality of our CENet framework. We apply BERT and XLNet as the language model of CENet, respectively. Meanwhile, we retrain some baselines using BERT and XLNet as the language model, respectively. Combining the Table I and Table II, we have some new observations as follows.

1) CENet(B) and CENet(X) still outperform these baselines under the condition of using the same language model, which further demonstrates the effectiveness of our CENet framework.

2) Results of some multimodal methods including TFN(B), LMF(B) and MulT(B) are inferior to those of single modality method T(B). When using XLNet as the language model, the situation is the same. However, in table I, the results of TFN, LMF and MulT are better than those of single modality method T(G) which uses GloVe word embeddings. This illustrates that BERT and XLNet pre-trained language models are not suitable for TFN, LMF and MulT. It is worth noting that GloVe is used in their respective original papers. We argue that the reason for this phenomenon is the three multimodal methods consider textual, visual and acoustic modalities to be equally important and deeply couple them. When using GloVe word embeddings, the performances of textual modality and nonverbal modalities are close. Therefore, the three multimodal methods can achieve improvements compared to unimodal methods. However, when the GloVe word embeddings are replaced by BERT or XLNet word embeddings, the performance of text modality is greatly improved and the balance between verbal and nonverbal modalities is broken. In this case, the deep-coupled fusion method may cause emotional information loss in text modality.

3) The performances of MISA(B) and Self-MM(B) are improved compared with T(B). However, the situation is changed when XLNet is used. Multimodal methods including MISA(X) and Self-MM(X) do not achieve improvement over unimodal method T(X). The same holds when SentiLARE is used as the language model. Both MISA and Self-MM utilize multitask framework to aid the learning of modality-specific and modality-invariant representations. In this case, text representations can maintain good independence when fusing with nonverbal representations. However, MISA and Self-MM treat text, acoustic and visual modalities as equally important. Such methods will encounter difficulties When the performance gap between verbal and nonverbal modalities widens further.

4) MAG-BERT(B) outperforms T(B) and MAG-XLNet(X) also performs better than T(X). The MAG series is similar to our CENet. All of them use nonverbal emotional information to enhance text representations. Such fusion methods can mitigate the problem that performances of verbal and nonverbal modalities are extremely imbalanced. We also observe that the performance of MAG-SentiLARE(S) is close to T(S). But it cannot surpass T(S) comprehensively. The enhancement strategy of MAG is to use nonverbal features which are aligned with words to enhance the text representation within a pre-trained language model. It does not have the ability to capture long-range emotional cues from nonverbal modalities. Furthermore, there are large distribution differences among modalities, while the nonverbal features used in MAG are not specially processed for this problem. Therefore, the enhancement information captured by MAG is not sufficient to boost the text representation within SentiLARE.

Our CENet utilizes emotional cues captured from long-range nonverbal features to enhance the text representation within a pre-trained language model. And a transformation strategy is used to reduce the distribution differences between heterogeneous modalities. Extensive experiments have shown that our CENet combined with several pre-trained language models can achieve good results.

TABLE III
ACC-2 ON MOSI BY VARYING CLUSTER NUMBER. V AND A DENOTE VISUAL AND ACOUSTIC MODALITIES, RESPECTIVELY

Clusters	8	16	32	64	128
V	58.99	61.59	59.76	58.38	58.08
A	57.62	60.67	60.06	58.97	58.38

H. Effect of Cluster Number

In this paper, we design a clustering-based feature transformation strategy which applies vector quantization for acoustic and visual features. The cluster number is an important factor for feature transformation. It determines whether the index vector adequately represents the original features. To study the effect of cluster number, we apply Transformer to extract non-verbal representations and use two fully connected layers to predict sentiment intensity on the MOSI dataset. From Table III, we can see that 16 is an appropriate cluster number. This implies that fewer clusters are unable to separate differential frames completely, while more clusters split similar frames.

I. Feature Transformation Vs. Raw Feature

To validate the role of our feature transformation strategy, we design two comparative experiments.

TABLE IV
RESULTS OF DIFFERENT TYPES OF INPUT. V AND A DENOTE VISUAL AND ACOUSTIC RESPECTIVELY. RAW FEATURES AND NONVERBAL FEATURE EMBEDDING ARE REPRESENTED BY "RAW" AND "EMBEDDING" RESPECTIVELY

Modality	Input	Acc-2	F1	MAE	Corr
V	Raw	58.54	58.79	1.462	0.209
	Embedding	61.59	58.97	1.355	0.217
A	Raw	57.32	57.48	1.444	0.172
	Embedding	60.67	58.82	1.403	0.179

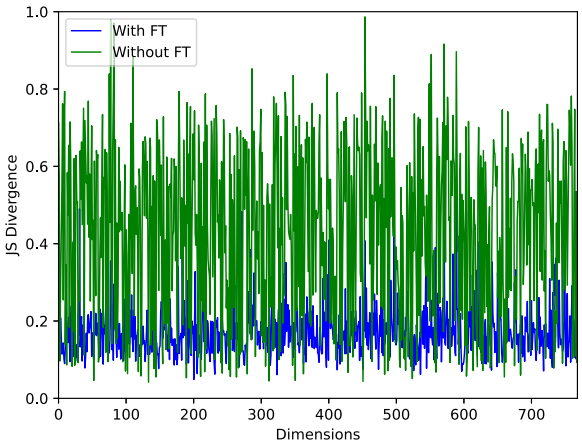

Fig. 6. A diagram of JS divergences between the text representations and the nonverbal enhancement embeddings with and without the feature transformation (FT) strategy.

In the first experiment, we compare the performance of a two-layer Transformer with the input of transformed feature embeddings and raw hand-crafted features, respectively. The transformed feature embeddings are obtained by combining the feature transformation strategy with the embedding layer. A fully connected layer is utilized as the head to predict sentiment intensity. As shown in Table IV, the transformed feature embeddings outperform raw features on both visual and acoustic modalities. This suggests that the nonverbal vocabularies can replace the nonverbal raw features without performance loss.

In the second experiment, we compare the Jensen-Shannon (JS) divergences of feature distribution between text representations and nonverbal enhancement embeddings with and without using the feature transformation strategy, respectively. Specifically, we train two CENet models using the transformed feature embeddings and the raw features on the validation set of the MOSI dataset, respectively. Then we can get two kinds of nonverbal enhancement embedding and corresponding text representation which are obtained in the CE module. After that, we standardize the collected nonverbal enhancement embeddings and corresponding text representations and then compare their distribution similarity of each dimension (768 dimensions in total) by calculating the JS divergence. Fig. 6 presents the JS divergence between the text representations and the nonverbal enhancement embeddings with and without feature transformation strategy, respectively. We can observe that the average JS divergence using the feature transformation strategy is much smaller than that using the raw features. This suggests that the feature transformation strategy can reduce the distribution differences between verbal and nonverbal features, thus facilitating the emotional information in nonverbal data to be integrated into text representations.

In the ablation study section, we will further explore the performance impact of applying the feature transformation strategy to CENet.

TABLE V
RESULTS OF EMBEDDING THE CE MODULE AT DIFFERENT LAYERS

Layer	Acc-2	F1	MAE	Corr
0_{th}	90.09	90.03	0.588	0.865
1_{th}	90.85	90.82	0.570	0.870
2_{th}	90.09	90.08	0.598	0.870
4_{th}	89.63	89.59	0.600	0.865
8_{th}	89.63	89.59	0.617	0.860
10_{th}	89.48	89.50	0.610	0.869
12_{th}	89.48	89.41	0.602	0.869

J. Effects of Embedding the CE Module At Different Layers

The effects of embedding the CE module at different layers on the performance of CENet are studied here. We embed the CE module behind the $i \in \{0, 1, 2, 4, 6, 8, 10, 12\}$-th layer of SentiLARE and record the corresponding results on the MOSI dataset in Table V, where the 0-th layer represents the embedding layer. It is clear that embedding the CE module at a lower layer can achieve better performance than embedding it at a higher layer. For text modality, the representations are more high-level at higher layers. As the non-verbal representations are relatively low-level features, embedding the CE module at a lower layer will be more appropriate.

K. Ablation Study

To further study the influence of each component in CENet, we perform a comprehensive ablation analysis on the MOSI dataset. The followings are some variants of CENet.

- $A \rightarrow T$: Let language be the target modality and acoustic be the source modality. We first use cross-modal attention to adapt acoustic modality to language modality, afterward input the obtained text-based acoustic embedding to the remaining layers of SentiLARE. The adaptation process is performed after the first layer in SentiLARE.
- $V \rightarrow T$: Similar to settings in $A \rightarrow T$, only replace acoustic modality with visual modality.
- T: CENet without integrating nonverbal information, which is equivalent to SentiLARE.
- $T, A \rightarrow T$: CENet without integrating visual information.
- $T, V \rightarrow T$: CENet without integrating acoustic information.
- $T, A \rightarrow T, V \rightarrow T$: Typical CENet.
- $T, A_{raw} \rightarrow T, V_{raw} \rightarrow T$: CENet without using the feature transformation strategy.
- T, A_{raw}, V_{raw}: Using MAG instead of CE module. The feature transformation is also not used.

Table VI shows the results of all the variants of CENet. We can draw the following observations.

TABLE VI
RESULTS OF ABLATION STUDIES ON MOSI DATASET. * MEANS $p < 0.05$ UNDER MCNEMAR'S TEST FOR BINARY CLASSIFICATION. HERE, THE STATISTICAL SIGNIFICANCE TESTS ARE COMPARED WITH $\{T\}$, $\{T, A_{raw} \rightarrow T, V_{raw} \rightarrow T\}$ AND $\{T, A_{raw}, V_{raw}\}$

Model	Acc-2	F1	MAE	Corr
$A \rightarrow T$	78.81	78.99	0.975	0.638
$V \rightarrow T$	76.98	76.82	1.032	0.595
T	89.94	89.88	0.580	0.864
$T, A \rightarrow T$	90.55	90.51	0.599	0.865
$T, V \rightarrow T$	90.40	90.37	0.588	0.869
$T, A \rightarrow T, V \rightarrow T$	**90.85***	**90.82**	**0.570**	**0.870**
$T, A_{raw} \rightarrow T, V_{raw} \rightarrow T$	90.40	90.37	0.589	0.863
T, A_{raw}, V_{raw}	89.62	89.57	0.573	0.867

- Comparing the results of $A \rightarrow T$ and $V \rightarrow T$ with T, we find that the performance of text representations outperforms that of text-based nonverbal embeddings. This is plausible for two reasons: 1) The pre-trained language model is not adaptable to nonverbal features. 2) The features of nonverbal modality are inferior to that of text modality.
- $T, A \rightarrow T$ and $T, V \rightarrow T$ performs better than T. This indicates that both acoustic and visual modalities can provide useful information to enhance language representations by using the CE module.
- $T, A \rightarrow T, V \rightarrow T$ performs best. This indicates that combining the information of acoustic and visual modalities can further enhance language representations.
- $T, A_{raw} \rightarrow T, V_{raw} \rightarrow T$ outperforms T, A_{raw}, V_{raw} and T on Acc-2 and F1 score which demonstrates the effectiveness of the CE module. Moreover, the CE module enhances text representations using unaligned nonverbal data thus avoiding cumbersome engineering of alignment.
- Performance of $T, A \rightarrow T, V \rightarrow T$ is superior to $T, A_{raw} \rightarrow T, V_{raw} \rightarrow T$, which shows that the feature transformation strategy can indeed facilitate integrating nonverbal emotional information into text representations.

L. Qualitative Analysis

To clarify how CENet works, we show some cases where CENet predicts sentiment intensity with and without using the cross-modal enhancement embedding. Table VII presents some examples of how CENet adjusts sentiment intensity by incorporating the nonverbal information. In the first and third examples, CENet without cross-modal enhancement embeddings predicts sentiment polarity only with text modality, we find that the intensity of predicted sentiment is insufficient. When incorporating the enhancement embeddings, the predicted value is almost increased to the true sentiment intensity. In the second example, it is ambiguous to judge the sentiment polarity only from the text information. In this case, non-verbal information can help the model to determine the polarity of sentiment. These observations demonstrate that CENet can successfully utilize the information in acoustic and visual modalities to enhance sentiment prediction.

TABLE VII
EXAMPLES OF CENET WITH AND WITHOUT USING NONVERBAL ENHANCEMENT EMBEDDING. CENET(S) AND T(S) DENOTE THE NONVERBAL ENHANCEMENT EMBEDDING IS USED AND NOT USED, RESPECTIVELY

	Example	Label	CENet(S)	T(S)
T A V	It reached the new height of horrendous. emphasize serious expression	-2.4	-2.35	-1.59
T A V	He was the only character that slightly interesting. hesitant tone thinking and shaking head.	-0.8	-0.84	-0.09
T A V	Oh my gosh bad movie. excited voice smile and surprised	-2.8	-2.78	-1.88

V. CONCLUSION

In this paper, a cross-modal enhancement network (CENet) model is proposed to enhance the text representation within a pre-trained language model by utilizing the long-range visual and acoustic emotional information. In addition, a feature transformation strategy is introduced to reduce the distribution differences between verbal and nonverbal initial representations. Extensive experiments demonstrate the superior performance of CENet over state-of-the-arts on the benchmark datasets MOSI and MOSEI. In future work, we will design a fully end-to-end multimodal learning model to explore the interactions between verbal and nonverbal features.

REFERENCES

[1] A. Zadeh, P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[2] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
[4] Z. Yang *et al.*, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5753–5763.
[5] Y. Liu *et al.*, "RoBERta: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[6] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. 34th AAAI Conf. Artif. Intell.*, vol. 2020, pp. 8992–8999.
[7] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.
[8] W. Rahman *et al.*, "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.
[9] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "Covarep: A collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 960–964.
[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] Y.-H. Tsai *et al.*, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[12] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, 2004, pp. 271–278.
[13] S. Alhojely, "Sentiment analysis and opinion mining: A survey," *Int. J. Comput. Appl.*, vol. 2, no. 6, pp. 22–25, 2016.
[14] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.
[15] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. 2002 Conf. Empirical Methods Natural Lang. Process.*, 2002, pp. 79–86.
[16] X. Zhang and X. Zheng, "Comparison of text sentiment analysis based on machine learning," in *Proc. 15th Proc. Int. Symp. Parallel Distrib. Comput.*, 2016, pp. 230–233.
[17] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.
[18] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 720–728.
[19] R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.
[20] Y. Chen, J. Yuan, Q. You, and J. Luo, "Twitter sentiment analysis via Bi-sense emoji embedding and attention-based LSTM," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 117–125.
[21] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[22] H. Xu, B. Liu, L. Shu, and P. S. Yu, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 2324–2335.
[23] X. Hu, L. Bing, S. Lei, and P. Y. S., "Dombert: Domain-oriented language model for aspect-based sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1725–1731.
[24] D. Yin, T. Meng, and K. Chang, "Sentibert: A transferable transformer-based architecture for compositional sentiment semantics," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3695–3706.
[25] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "Sentilare: Sentiment aware language representation learning with linguistic knowledge," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 6975–6988.
[26] S. Jindal and S. Singh, "Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning," in *Proc. Int. Conf. Inf. Process.*, 2015, pp. 447–451.
[27] V. Campos, A. Salvador, X. Giró-i-Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction," in *Proc. 1st Int. Workshop Affect Sentiment Multimedia*, 2015, pp. 57–62.
[28] V. Campos, B. Jou, and X. Giró-i-Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vis. Comput.*, vol. 65, pp. 15–22, 2017.
[29] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.
[30] J. Yang *et al.*, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2513–2525, Sep. 2018.
[31] W. Chen and R. W. Picard, "Predicting perceived emotions in animated gifs with 3D convolutional neural networks," in *Proc. IEEE Int. Symp. Multimedia*, 2016, pp. 367–368.
[32] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 26–33.
[33] P. Tzirakis, J. Chen, S. Zafeiriou, and B. W. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inf. Fusion*, vol. 68, pp. 46–53, 2021.
[34] M. Wöllmer *et al.*, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 597–600.
[35] B. Schuller *et al.*, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 2253–2256.
[36] M. Neumann and T. N. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1263–1267.
[37] M. Xu, F. Zhang, and U. S. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *Proc. 10th Annu. Comput. Commun. Workshop Conf.*, 2020, pp. 1058–1064.
[38] P. Tzirakis, J. Zhang, and W. B. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5089–5093.
[39] Y. Li, T. Zhao, and T. Kawahara, "Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 2803–2807.
[40] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[41] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2539–2544.
[42] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. 16th Int. Conf. Data Mining*, 2016, pp. 439–448.
[43] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[44] Z. Liu *et al.*, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.
[45] Y. Wang *et al.*, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.
[46] A. Zadeh *et al.*, "Factorized multimodal transformer for multimodal sequential learning," 2019, *arXiv:1911.09826*.
[47] K. Yang, H. Xu, and K. Gao, "CM-BERT: Cross-modal BERT for text-audio sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 521–528.
[48] E. M. Peters *et al.*, "Deep contextualized word representations," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 2227–2237.
[49] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI*, Tech. Rep., 2018.
[50] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7464–7473.
[51] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2003, pp. 252–259.
[52] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiwordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. 7th Int. Conf. Lang. Resour. Eval.*, 2010, pp. 2200–2204.
[53] J. Cheong, T. Xie, S. Byrne, and L. Chang, "Py-feat: Python facial expression analysis toolbox," 2021, *arXiv:2104.03509*.
[54] A. Zadeh *et al.*, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.

Di Wang (Member, IEEE) 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. She has authored or coauthored several scientific articles in refereed journals including the IEEE Transactions on Pattern Analysis and Machine Intelligence, TIP, TMM, TCYB and TCSVT, and conferences including the SIGIR and IJCAI. Her research interests include machine learning and multimedia information retrieval.

Shuai Liu received the B.S. degree in computer science and technology in 2020 from Xidian University, Xi'an, China, where he is currently working toward the M.S. degree with the Guangzhou Institute of Technology. His research interests include machine learning and computer vision.

Quan Wang received the B.Sc., M.Sc., and Ph.D. degrees in computer science and technology from Xidian University, Xi'an, China. He is currently a Professor with the School of Computer Science and Technology, Xidian University. His research interests include input and output technologies and systems, image processing, and image understanding.

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

Lihuo He received the B.Sc. degree in electronic and information engineering and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2008 and 2013 respectively. He is currently an Associate Professor with Xidian University. His research interests include image/video quality assessment, cognitive computing, and computational vision.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Postdoctoral Research Fellow with the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, China, a Professor of pattern recognition and intelligent system, and a Professor of computer science and technology with the Chongqing University of Posts and Telecommunications, Chongqing, China. He has authored or coauthored six books and around 300 technical articles in refereed journals and proceedings. His research interests include image processing, computer vision, multimedia analysis, machine learning and pattern recognition. Prof. Gao is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier). He was the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a fellow of the Institute of Engineering and Technology and a fellow of the Chinese Institute of Electronics.