2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650083

Learning Modality-Complementary and Eliminating-Redundancy Representations with Multi-Task Learning for Multimodal Sentiment Analysis

Xiaowei Zhao, Xinyu Miao, Xiujuan Xu*, Yu Liu, Yifei Cao
School of Software Technology, Dalian University of Technology
{xjxu, xiaowei.zhao, yuliu}@dlut.edu.cn, {mxy0204, yfcao}@mail.dlut.edu.cn

***Abstract*—A crucial issue in multimodal language processing is representation learning. Previous works joint training the multimodal and unimodal tasks to learn the consistency and difference of modality representations. However, due to the lack of cross-modal interaction, the extraction of complementary features between modalities is not sufficient. Moreover, during multimodal fusion, the generated multimodal embeddings may be redundant, and unimodal representations also contain noise information, which negatively influence the final sentiment prediction. To this end, we construct a Modality-Complementary and Eliminating-Redundancy multi-task learning model (MCER), and additionally add a cross-modal task to learn complementary features between two modal pairs through gated transformer. Then use two label generation modules to learn modality-specific and modality-complementary representations. Additionally, we introduce the multimodal information bottleneck (MIB) in both multimodal and unimodal tasks to filter out noise information in unimodal representations as well as learn powerful and sufficient multimodal embeddings that is free of redundancy. Last, we conduct extensive experiments on two popular sentiment analysis benchmarks, MOSI and MOSEI. Experimental results demonstrate that our model significantly outperforms the current strong baselines.**

***Index Terms*—multimodal sentiment analysis, cross-modal transformer, multimodal information bottleneck, multi-task learning, label generation**

I. Introduction

In recent years, with the unprecedented development of short video platform, more and more people are sharing their lives and expressing their opinions in the form of videos on social media [1]–[3]. Multimodal sentiment analysis (MSA) attracts increasing attention [4], [5]. Such video contains data in three modalities: visual, audio and text. Different modalities in the same data segment are often complementary [6]. On the other hand, different modalities often possess unique statistical properties that make them somewhat independent of each other [7]. Therefore, a key issue in multimodal language processing is how to extract and integrate meaningful information to obtain effective modality representations and effectively integrate heterogeneous data.

Some previous methods in MSA focus on developing complex fusion mechanisms to learn efficient multimodal embeddings. For example, tensor fusion [8]–[10] and attention-based fusion [4], [11], these fusion methods are effective, but they have high complexity and the learned high-dimensional multimodal embeddings are inevitably redundant. Furthermore, unimodal representations often contain noise information, especially non-lexical modalities like visual and audio [12], [13], which may negatively affect the final prediction. To alleviate these issues, inspired by [14], we introduce multimodal information bottleneck (MIB) based on mutual information (MI) in multimodal task and unimodal task. The MIB aims to maximize the MI between the encoded representation and the labels, while minimizing the MI between the encoded representation and the input.

Since the complementary information between different modalities cannot be sufficiently learned only through multimodal and unimodal task, so we introduce a cross-modal task. Consistent with previous research in this field [4], [8], we adopted ternary-symmetric architectures, in which the bidirectional relationship in each modal pair is modeled with gated transformer [15].

Due to the unified multimodal annotation, existing methods are restricted in capturing differentiated and complementary information. Therefore, inspired by [16], we introduce label generation modules based on multimodal labels and modality representations in unimodal and cross-modal tasks. They respectively generate labels specific to a single modality and multimodal labels that pay more attention to a certain modality.

To summarize, the main contributions of our work are as follows:

- We propose a multi-task learning model, which includes a multimodal task, a unimodal task and a cross-modal task. We joint training the above three tasks to learn the consistency, complementary and difference of modality representations.
- We introduce multimodal information bottleneck (MIB), aiming to learn powerful and sufficient multimodal embeddings that is free of redundancy and filter out the noise information to learn effective unimodal representations.
- We introduce label generation modules based on multimodal labels and modality representations, in order to capture differentiated and complementary information in different modalities.

- Extensive experiments on two popular MSA benchmark datasets demonstrate that MCER gains superior or comparable results to the current strong baselines.

II. Related Work

In this section, we briefly overview some related work in the domain of multimodal sentiment analysis, information bottleneck and multi-task learning.

A. Multimodal Sentiment Analysis

Multimodal sentiment analysis is an important task in NLP, which focus on tackling acoustic, visual, and textual information to comprehend varied human emotions [17].

For multimodal fusion methods, previous more advanced research efforts focus on developing complex fusion strategies. Previously, tensor-based fusion and its low-rank variants has received much attention [18]–[20]. For example, [8] proposed a Tensor Fusion Network that adopts outer product to capture interaction features between modalities. Meanwhile, graph-based fusion methods have emerged [21]–[23], which can effectively extract interaction across time series. For example, [23] proposed Graph-MFN regards each modality as one node and uses a dynamic fusion graph to fuse features. More recently, fusion methods based on mutual information have emerged. For example, [24] proposed a hierarchical MI maximization framework to reduce the loss of valuable task-related information.

For representation learning methods, [25] proposed the Recurrent Attended Variation embeddings Network, which dynamically shifts the word representations based on nonverbal cues. Subsequently, [26] proposed a simple and flexible multimodal learning framework that learns modality-invariant and modality-specific representations. Recently, [16] proposed SELF-MM to learn modality-specific representations with self-supervised unimodal label generation module. More recently, [27] mixed features on three axes, the distinct multimodal information is effectively transmitted and shared during the mixing process to extract important features.

Different from previous studies, we adopt a simple concatenation strategy to fuse different modalities and adopt MIB to learn minimal and sufficient modality representations and multimodal embeddings.

B. Information bottleneck

Information bottleneck (IB) is a method in information theory. IB is based on mutual information aiming to learn the minimal sufficient representations [28]. IB originally proposed to process signal [29], and it cannot be applied to deep learning. Later, the emergence of variational information bottleneck (VIB) bridges the above gap [30]. [31], [32] first adopted VIB to constrain the information flow of each view, which extend VIB for multi-view learning. Meanwhile, [33] developed a deep multi-view IB theory. Nowadays, IB and VIB are widely used in natural language processing [34], reinforcement learning [35] and computer vision [36]. More recently, [14] proposed Multimodal Information Bottleneck (MIB) simultaneously applies the IB principle to unimodal representation learning and complicated multimodal representation learning, which enables the learning of more discriminative and expressive features.

Inspired by [14], we introduce MIB into our work that manage to reduce the redundancy of the generated multimodal embeddings and filter out noise information of each unimodal representation.

C. Multi-task Learning

The purpose of multi-task learning is to leverage shared information across multiple related tasks to help improve generalization performance across all tasks [37]. In current multimodal sentiment analysis research, most multi-task learning methods adopt hard-sharing method, different tasks share low-level features and parameters. Recently, multi-task learning is wildly applied in MSA. For example, [38] leverage the inter-dependence of two related tasks (i.e. sentiment and emotion) in improving each others performance using an effective multimodal framework. [39] proposed a multi-task learning framework based on late fusion. More recently, [16] joint training the multimodal and unimodal tasks to learn the consistency and difference.

In our work, MCER adopt the hard sharing strategy to learn modality-complementary and modality-specific representations.

III. Methodology

In this section, we explain multi-task learning model in detail.

A. Problem Definition

In the MSA task, the input to the model includes three modalities: text(t), audio(a) and visual(v), these inputs are unimodal raw sequences extracted from the same video clip, which can denoted by $U_t \in \mathbb{R}^{T_t \times d_t}$, $U_a \in \mathbb{R}^{T_a \times d_a}$, $U_v \in \mathbb{R}^{T_v \times d_v}$, where T_m and d_m represent sequence length and feature vector size of modality m respectively, where $m \in \{t, a, v\}$.

B. Overall Architecture

Figure 1 shows the overall architecture of MCER. We first obtain low-level unimodal features with the corresponding feature extractors, then the three tasks share the bottom low-level unimodal features. Next, in the multimodal task, multimodal embeddings is generated for the two label generation modules and obtain the final predictions. In the cross-modal task, complementary modal representations are learned through the gated transformer to obtain the final cross-modal predictions and modality-complementary labels(c-labels) that generated through Complementary Label Generation Module (CLGM). In the unimodal task, learning unimodal representations to generate unimodal predictions and obtain modality-specific labels(s-labels) by Unimodal Label Generation Module (SLGM).

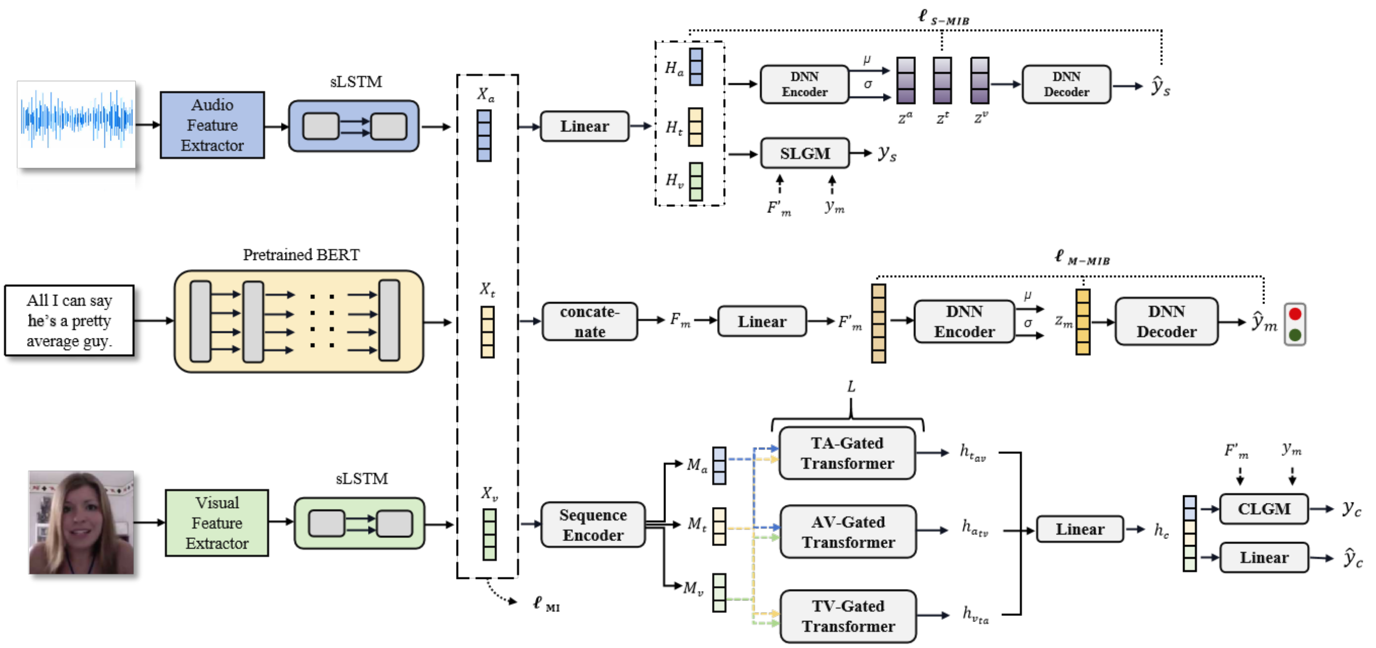

Fig. 1. The overall architecture of MCER. The low-level features of the three modalities are extracted separately, and then the three tasks share the low-level modal features to generate high-level modality representations. $\hat{y}_s, \hat{y}_m$ and $\hat{y}_c$ are the predicted values of unimodal, multimodal and cross-modal tasks respectively. y_m is the label annotated by humans, y_s and y_c are labels generated by unimodal and cross-modal tasks.

C. MCER

In this section, we will introduce the three tasks in detail.

Unimodal Encoding. Specifically, in the text modality, we use the pretrained BERT [40] to encode the input sentence and extract the head embeddings from the last layer's output as X_t.

$$X_t = \mathrm{BERT}(U_t; \theta_t^{bert}) \tag{1}$$

For visual and audio, following previous works [16], we employ a unidirectional Long Short-Term Memory (sLSTM) [41] to capture the temporal features of these modalities.

$$X_m = \mathrm{sLSTM}(U_m; \theta_m^{lstm}) \quad m \in \{a, v\} \tag{2}$$

Next, we calculate the mutual information [24] between two unimodal representations and use it as the loss $\mathcal{L}_{MI}$ to initially remove the noise information in the unimodal representations and retain as many task-related features as possible to improve the performance of subsequent tasks.

Multimodal Task. In multimodal task, we first concatenate all unimodal representations and project them into a lower-dimensional space.

$$F_m^{'} = \mathrm{ReLU}(W_1^m[X_t; X_a; X_v] + b_1^m) \tag{3}$$

where ReLU is the relu activation function.

Last, the fusion representation $F_m^{'}$ is input into DNN encoder and DNN decoder to obtain the final prediction.

$$z_m = \mathrm{DNNEncoder}(W_2^m F_m^{'} + b_2^m) \tag{4}$$

$$\hat{y}_m = \mathrm{DNNDecoder}(W_3^m z_m + b_3^m) \tag{5}$$

Unimodal task. The unimodal task first projects the shared low-level unimodal features into a new feature space to reduce the dimensional difference between different modalities.

$$H_s = \mathrm{ReLU}(W_1^s X_s + b_1^s) \tag{6}$$

Then the modality representations H_s is input into DNN encoder and DNN decoder to obtain the final prediction.

$$z^s = \mathrm{DNNEncoder}(W_2^s H_s + b_2^s) \tag{7}$$

$$\hat{y}_s = \mathrm{DNNDecoder}(W_3^s z^s + b_3^s) \tag{8}$$

where $s \in \{t, a, v\}$.

Then we get s-labels through SLGM. Details of the SLGM are discussed in Section 3.5.

$$y_s = \mathrm{SLGM}(y_m, F_m^{'}, H_s) \tag{9}$$

where y_m (m-labels) is the human-annotated multimodal labels.

In the training stage, we joint learn the multimodal task, unimodal task and cross-modal task under m-labels, s-labels and c-labels supervision. And we only use the output $\hat{y}_m$ of the multimodal task as the final test stage output.

Cross-modal task. We first use a single-layer Bidirectional Gated Recurrent Unit (BiGRU) [42] and a linear projection layer to encode shared low-level unimodal features and convert all hidden vectors to the same length to facilitate further processing.

$$M_s = \mathrm{SequenceEncoder}(X_s; \theta^{SE}) \tag{10}$$

These outputs serve as the initial inputs to the L stacked gated transformers. L stacked gated transformers is constructed for every two modality pairs to form a ternary symmetric structure. In this structure, the modality representation pairs can complement the missing cues with their counterparts. For the implementation of the gated transformer, we refer to the work of [15].

In order to enhance control over information flow, we introduce two gates in gated transformer: the filter gate g_f, which decides the proportion of the current modality's components to be kept forwarding, and the complementary gate g_c, which decides the proportion of auxiliary modality to be injected to the current modality.

We adopt a bidirectional GRU and average pooling to the output of $(i-1)_{th}$ layer to acquire sequence-level hidden representations:

$$x_s^i = \text{avgpool}(\text{BiGRU}(M_s^{i-1};\theta_s^i)) \tag{11}$$

Then the obtained sequence-level hidden representations input into the i_{th} layer two gates:

$$g_f^i = \sigma(W_f^{i\ s}[x_s^i||x_m^i]) \tag{12}$$

$$g_c^i = \sigma(W_c^{i\ m}[x_m^i||x_s^i]) \tag{13}$$

where $m \in \{t, a.v\}$, but $s \neq m$, $W \in \mathbb{R}^{2d\times d}$ is the projection matrix, $||$ represents concatenation and σ is the sigmoid activation function.

The generated two gate signals employed on the output of multi-head attention, then the attention results o^i pass through the feed-forward network to produce the final output of the current layer:

$$o^i = \text{MHA}(W_Q^i x_s^i, W_K^i x_m^i, W_V^i x_m^i) \tag{14}$$

$$\bar{M_s^i} = \text{LN}(g_f^i \odot M_s^i + g_c^i \odot o^i) \tag{15}$$

$$M_s^i = \text{LN}(\bar{M_s^i} + \text{FFN}(\bar{M_s^i})) \tag{16}$$

where s and m can be replaced with each other, but $s \neq m$. MHA represents multi-head attention, $\odot$ means componentwise multiplication and LN is layer normalization.

The last layer of each gated transformer has two outputs, so there are six outputs in total: H_{ta_t}, H_{ta_a}, H_{tv_t}, H_{tv_v}, H_{av_a}, H_{av_v}. We first extract the heads representations from them and then concatenate two heads representations that from the same gated transformer to get: H_{ta}, H_{tv}, H_{av}, which have learned complementary information from each other.

Next, we concatenate the above representations two by two to get $h_{t_{av}}, h_{a_{tv}}, h_{v_{ta}}$, then input them into liner layer to get multimodal embeddings that pay more attention to a certain modality, the resulting representation is used to generate the final prediction:

$$h_c = \text{ReLU}(W_c^1 h_p + b_c^1) \tag{17}$$

$$\hat{y}_c = W_c^2 h_c + b_c^2 \tag{18}$$

where $p \in \{t_{av}, a_{tv}, v_{ta}\}$ and $c \in \{m_t, m_a, m_v\}$. The resulting multimodal representations h_c, such as h_{m_t}, which is obtained from $h_{t_{av}}$. In $h_{t_{av}}$, both A and V have learned the supplementary information in T, so the multimodal representation h_{m_t} pay more attention to text modality.

Then we adopt CLGM to generate c-labels. Details of the CLGM are discussed in Section 3.5.

$$y_c = \text{CLGM}(y_m, F_m^{'}, h_c) \tag{19}$$

Last, we calculate the mean absolute error (MAE) between prediction $\hat{y}_c$ and c-labels y_c as the task loss:

$$\mathcal{L}_{c-task} = \text{MAE}(\hat{y}_c, y_c) \tag{20}$$

In order to maintain the mutual independence between different modalities in the same complementary module, inspired by [15], for each gated transformer we use the regularization effect produced by the discriminator loss to distinguish different modalities. We calculate the Binary Cross Entropy between predictions and their corresponding pseudo labels that are automatically generated during training time as the discriminator loss:

$$\mathcal{L}_{dc}^i = -\frac{k}{2n}\sum_{r=1}^{2n/k}(c_r^i log\hat{c}_r^i + (1-c_r^i)log(1-\hat{c}_r^i)) \tag{21}$$

where k is the group size for grouping modality representations, r represents the r_{th} group, c_r^i and $\hat{c}_r^i$ is the representation obtained after grouping, n is the batch size and i represents the i_{th} sample.

Because the discriminator loss is a layer-wise loss and the complementary module of each modality pair consists of L gated transformers, so we sum up the results that are computed in each gated transformer:

$$\mathcal{L}_{dc} = \frac{1}{n}\sum_{j=1}^{n}(\frac{\lambda K}{2L}\sum_{i=1}^{L}\sum_{M}\mathcal{L}_{dc}^{M,i}) \tag{22}$$

where $M \in \{TA, TV, AV\}$ and λ is a tunable parameter to control the power of regularization.

D. MIB

In multimodal task, all information from the three modalities be preserved through a simple concatenate strategy. Motivated by [14], so we adopt MIB in multimodal task, denoted by M-MIB to regularize multimodal representation, the purpose is to remove redundancy in the multimodal embedding $F_m^{'}$. The objective function of M-MIB can be defined as:

$$L_{M-MIB} = I(y_m; z_m) - \beta I(F_m^{'}; z_m) \tag{23}$$

We optimize L_{M-MIB} by maximizing the lower bound of L_{M-MIB}, denoted by $\mathcal{L}_{m-task}$. Maximizing the lower bound can be implement in the following two parts.

The first part is to maximize the mutual information between the target y_m and the compressed encoded representation z_m, which aims to encourage z_m to be maximally predictive of the target y_m. It can be approximated by Monte Carlo sampling [43]:

$$\mathcal{L}_{m-task} = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{\epsilon\sim p(\epsilon)}[\log\ q(y_i \mid z_i)]\right] \tag{24}$$

TABLE I

RESULTS ON CMU-MOSI AND CMU-MOSEI. IN ACC-2 AND F1, THE LEFT OF THE "/" CORRESPONDS TO "NEGATIVE/NON-NEGATIVE" AND THE RIGHT CORRESPONDS TO "NEGATIVE/POSITIVE".†: FROM ([24]), ‡: FROM ([27]). BEST RESULTS ARE MARKED IN BOLD.

Models	CMU-MOSI					CMU-MOSEI				
	MAE	Corr	Acc-7	Acc-2	F1	MAE	Corr	Acc-7	Acc-2	F1
TFN†	0.901	0.698	34.9	-/80.8	-/80.7	0.593	0.700	50.2	-/82.5	-/82.1
LMF†	0.917	0.695	33.2	-/82.5	-/82.4	0.623	0.677	48.0	-/82.0	-/82.1
MFM†	0.877	0.706	35.4	-/81.7	-/81.6	0.568	0.717	51.3	-/84.4	-/84.3
ICCN†	0.862	0.714	39.0	-/83.0	-/83.0	0.565	0.713	51.6	-/84.2	-/82.2
MulT†	0.861	0.711	-	81.5/84.1	80.6/83.9	0.580	0.703	-	-/82.5	-/82.3
MISA†	0.804	0.764	-	80.79/82.10	80.77/82.03	0.568	0.724	-	82.59/84.23	82.67/83.97
MAG-BERT†	0.727	0.781	43.62	82.37/84.43	82.50/84.61	0.543	0.755	52.67	82.51/84.82	82.77/84.71
SELF-MM†	0.712	0.795	**45.79**	82.54/84.77	82.68/84.91	**0.529**	0.767	53.46	82.68/84.96	82.95/84.93
CubeMLP‡	0.770	0.767	45.5	-/85.6	-/85.5	**0.529**	0.760	**54.9**	-/85.1	-/84.5
MCER	**0.699**	**0.799**	45.34	**84.69/86.74**	**84.63/86.73**	0.532	**0.769**	54.00	**83.49/86.10**	**83.80/86.03**

Moreover, or regression task, we formulate $\log\, q(y_i \mid z_i)$ as:

$$\begin{aligned} \log\, q(y_i \mid z_i) &= -\|\, y_i - D(z_i;\theta_d)\, \|_1 + C \\ &= -\|\, y_i - \hat{y}_i\, \|_1 + C \end{aligned} \tag{25}$$

where C is a constant, D is a decoder, and $\hat{y}_i$ is the prediction output by M-MIB. Under this situation, maximizing of $I(y_m; z_m)$ is converted to minimize the mean absolute error (MAE) between the prediction $\hat{y}_m$ and the target y_m.

The second part is to minimize the mutual information between the compressed encoded representation z_m and the multimodal embedding $F_m^{'}$, which enforce the z_m to only preserve the information in $F_m^{'}$ that is discriminative to the prediction. To use deep neural networks to optimize the above objectives, we use a standard normal Gaussian distribution and a Gaussian distribution with mean μ and variance σ to approximate [30], [33], and calculate the KL divergence between two Gaussian variables. It can be converted as:

$$\mathcal{L}_m = \frac{1}{n}\sum_{i=1}^{n}\Big[-\beta \cdot KL\Big(\mathcal{N}(\mu_{z_i}, \sigma_{z_i}) \mid\mid \mathcal{N}(0, I)\Big)\Big] \tag{26}$$

where n is batch size, and i is the subscript that indicates each sample. Under this situation, minimizing of $I(F_m^{'}; z_m)$ is converted to minimize $\mathcal{L}_m$.

Simultaneously, same principle as M-MIB, we adopt MIB in unimodal task denoted by S-MIB to filter out the noise information in the unimodal representations, the objective of S-MIB is:

$$L_{S-MIB} = \sum_{s}\Big[I(y_s; z^s) - \beta I(H_s; z^s)\Big] \tag{27}$$

L_{S-MIB} can be approximated by the following objective, respectively denoted by $\mathcal{L}_{s-task}$ and $\mathcal{L}_{avl}$:

$$\mathcal{L}_{s-task} = \frac{1}{n}\sum_{i=1}^{n}\sum_{s}\Big[\mathbb{E}_{\epsilon\sim p(\epsilon)}[\log q(y_i \mid z_i)]\Big] \tag{28}$$

$$\begin{aligned} \mathcal{L}_{avl} = \frac{1}{n}\sum_{i=1}^{n}\sum_{s}\Big[&-\beta \cdot KL\Big(\mathcal{N}(\mu_{z_i}, \sigma_{z_i}) \\ &\mid\mid \mathcal{N}(0, I)\Big)\Big] \end{aligned} \tag{29}$$

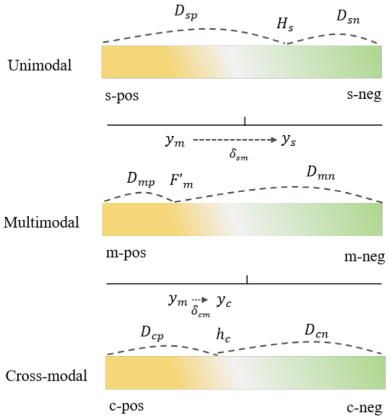

Fig. 2. Cross-modal label and unimodal label generation example. Multimodal representation $F_m^{'}$ is closer to the positive center while cross-modal representations h_c slightly closer to positive center and unimodal representation H_s is closer to the negative center . Therefore, cross-modal supervision y_c and unimodal supervision y_s are added a offset to the multimodal label y_m.

where $s \in \{t, a, v\}$, $\log\, q(y_i \mid z_i^s)$ is the mean absolute error (MAE) between the prediction $\hat{y}_s$ and the s-labels y_s.

E. LGM

We introduced two label generation modules: SLGM [16] aims to generate unimodal self-supervision labels in a unimodal task, while CLGM aims to generate multimodal self-supervision labels in a cross-modal task. Here, we take the cross-modal label generation module as an example to introduce. The CLGM is based on multimodal annotations and modality representations, which calculate the offset according to the relative distance from modality representations to class centers, as shown in Figure 2.

Therefore, inspired by [16], the self-supervised multimodal labels, dominated by a certain modality, can be obtained using the following formula:

$$y_c = y_m + \delta_{cm} \tag{30}$$

TABLE II
DATASET STATISTICS IN MOSI AND MOSEI.

Dateset	#Train	#Valid	#Test	#Total
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856

TABLE III
HYPERPARAMETERS FOR BEST PERFORMANCE.

Item	CMU-MOSI	CMU-MOSEI
batch size	32	32
learning_rate_bert	4.5e-5	2e-5
learning_audio	9e-4	9e-4
learning_rate_visual	6e-3	4.5e-3
learning_rate_other	7e-4	5.5e-4
α	1.9	1.8
β	0.001	0.45
γ	0.05	0.3
η	0.1	0.2

TABLE IV
ABLATION RESULTS OF THE EFFECTS OF DIFFERENT LOSS.

Models	MAE	Corr	Acc-7	Acc-2	F1
MCER	**0.699**	**0.799**	45.34	**84.69/86.74**	**84.63/86.73**
w/o $\mathcal{L}_m$	0.724	0.795	**46.06**	83.38/85.37	83.33/85.37
w/o $\mathcal{L}_{avl}$	0.721	0.789	45.63	82.51/84.91	82.27/84.77
w/o $\mathcal{L}_{dc}$	0.724	0.789	45.77	82.65/85.06	82.45/84.95
w/o $\mathcal{L}_{MI}$	0.720	0.792	45.19	83.09/84.91	83.06/84.93

TABLE V
RESULTS FOR MULTIMODAL SENTIMENT ANALYSIS WITH DIFFERENT TASKS USING MCER. M, S, C REPRESENT THREE TASKS, $m_t, m_a m_v$ REPRESENT CROSS-MODAL TASKS DOMINATED BY A CERTAIN MODALITY.

Models	MAE	Corr	Acc-7	Acc-2	F1
MCER	**0.699**	**0.799**	**45.34**	**84.69/86.74**	**84.63/86.73**
M, S	0.732	0.792	43.73	83.53/85.06	83.43/85.01
M, S, C_m$_t$	0.735	0.798	44.02	83.38/85.21	83.31/85.19
M, S, C_m$_a$	0.749	0.796	43.69	82.80/84.60	82.77/84.63
M, S, C_m$_v$	0.765	0.793	42.57	83.24/84.76	83.22/84.79
M, S, C_m$_t$_m_a	0.727	0.797	44.31	84.26/86.43	84.16/86.40
M, S, C_m$_t$_m_v	0.708	0.797	45.30	83.38/85.82	83.22/85.74
M, S, C_m$_a$_m_v	0.735	0.796	44.46	83.24/84.76	83.23/84.80

The $\delta_{cm} = \frac{\alpha_c - \alpha_m}{2} * \frac{y_m + \alpha_m}{\alpha_m}$, where α evaluates the relative distance from the modality representation to the positive center and the negative center.

However, the above generated c-labels are unstable due to the dynamic changes of modality representations. Therefore, we combine the newly generated value with the historical value to obtain the final c-labels.

$$y_c^{(i)} = \begin{cases} y_m & i = 1 \\ \frac{i-1}{i+1} y_c^{(i-1)} + \frac{2}{i+1} y_c^i & i > 1 \end{cases} \tag{31}$$

where $c \in \{m_t, m_a, m_v\}$. y_c^i is the new generated c-labels at the i_{th} epoch. $y_c^{(i)}$ is the final c-labels after the i_{th} epoch.

F. Training

Finally, we sum the losses of the three tasks to get the total task loss:

$$\mathcal{L}_{task} = \mathcal{L}_{m-task} + W_c \mathcal{L}_{c-task} + W_s \mathcal{L}_{s-task} \tag{32}$$

The weight $W_c = tanh(|y_c - y_m|)$ is the difference between c-labels and m-labels, the weight $W_s = tanh(|y_s - y_m|)$ is the difference between s-labels and m-labels. Therefore, the overall learning of the model is performed by minimizing:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha \mathcal{L}_m + \beta \mathcal{L}_{avl} + \gamma \mathcal{L}_{dc} + \eta \mathcal{L}_{MI} \tag{33}$$

where $\alpha, \beta, \gamma, \eta$ are the interaction weights that determine the contribution to the overall loss.

IV. EXPERIMENTS

In this section, we present some experimental details, including datasets, evaluation metrics, baselines, and experimental results.

A. Datasets and Metrics

We conduct experiments on two publicly available benchmark datasets in MSA: CMU-MOSI [44] and CMU-MOSEI [45]. Table II provide the split specifications of the two datasets.

Following the previous works [16], [24], [26], we used five different metrics to evaluate the performance of MCER: Pearson correlation (Corr) that measures the degree of prediction skew, mean absolute error (MAE), which calculates the error between predicted values and truth values, binary classification accuracy (Acc-2) and F1 score computed for positive/negative and non-negative/negative classification results, seven-class classification accuracy (Acc-7) which shows the proportion of predictions that correctly classified into the same interval of seven intervals between -3 and +3 as the corresponding truths.

B. Baselines

To validate the performance of MCER, we make a fair comparison with the following baselines.

TFN [8]. It calculates a multi-dimensional tensor by three-fold Cartesian product as fusion results.

LMF [18]. It decomposes high-order tensors into low rank factors to improve fusion efficiency.

MFM [46] It is a cycle style generative-discriminative model to learn the modality-specific generative features along with discriminative representations for classification.

ICCN [47] It learns Relationships between Text, Audio, and Video via Deep Canonical Correlation.

MulT [4]. It completes fusion process with directional pairwise cross-modal attention.

MISA [26]. It constructs two kinds of feature spaces to learn modality-invariant and modality-specific representation.

MAG-BERT [48]. It designs a multimodal adaptation gate and insert it into BERT backbone to refine the fusion process.

SELF-MM [16]. It learns the consistency and difference information of modalities through self-supervised multi-task model.

CubeMLP [27]. It proposes a multimodal feature processing framework based exclusively on MLP. CubeMLP mixes features on three axes: sequence, modality, and channel.

Experimental Settings. For visual and acoustic, we use LibROSA [49] and OpenFace2.0 [50], which both are prevalent tool kits for feature extraction and have been regularly employed before. We trained our model on a single RTX 3090 GPU and ran a random search for the best set of

TABLE VI

CASE STUDY FOR THE MCER ON MOSI. THE "M-LABELS" IS HUMAN-ANNOTATED, S-LSBELS(T/A/V) ARE MODALITY-SPECIFIC LABELS GENERATED BY SLGM, AND C-LABELS ARE MULTIMODAL LABELS DOMINATED BY A CERTAIN MODALITY THAT GENERATED BY CLGM.

	Text	Visual	Audio	m-labels	s-labes(t/a/v)	c-labels(m$_t/m_a/m_v$)
(A)	*You know I like kate hudson.*	Shake head	Normal volume Peaceful tone	1.2	1.27/-0.31/-0.29	0.91/0.80/0.82
(B)	*And um looking really forward* to seeing that movie.	Close eyes	Normal volume	1.8	1.57/-0.31/-0.30	1.27/1.11/1.20

hyper-parameters. We use Adam as the optimizer and hyper-parameters are given in Table III.

Results. We report the performance in Table I. In both datasets, we find that MCER yields better or comparable results to many baseline methods. MCER outperforms the current strong baseline models in Acc-2, F1 and Corr on CMU-MOSEI as well as all metrics except Acc-7 on CMU-MOSI. These outcomes preliminarily demonstrate the efficacy of our method in MSA tasks.

C. Ablation Study

To further examine the functionality of the MCER and the components introduced in this work, we conducted ablation studies on CMU-MOSI dataset.

We first investigate the contribution of each loss by removing it from our model. We conduct a series of ablation experiments as shown in Table IV. Through the results, it was found that removing any loss from the the model significant drop on performance, which proved the effectiveness of these component.

To show the benefits of the cross-modal task that we additionally introduce, we compare the effectiveness of combining different tasks. The results are shown in Table V. From the results, we can see that the introduce of cross-modal task can significantly improve model performance. Moreover, we found that tasks related to text modality are more helpful than visual and audio modalities.

D. Case Study

In order to further show the reasonability of the s-labels and c-labels, we selected two multimodal examples from the MOSI dataset, as shown in Table VI. In case A and B, m-labels are all positive, while single modalities show negative sentiments, so s-labels get negative offsets on m-labels. In addition, it can be seen from c-labels that the text-dominated multimodal labels is closer to the human-annotated labels than the other two, which also proves the dominance of the text modality. Therefore, the auto-generated s-labels and c-labels are significant, and they can aid in learning modality-specific and modality-complementary representations.

V. CONCLUSION

In this paper, we propose a multi-task learning model that learns modality-specific and modality-complementary representations by two label generation modules. MCER learns minimal and sufficient multimodal embeddings and unimodal representations with the introducing MIB. Extensive experiments validate the reliability and efficacy of the MCER and corresponding components.

ACKNOWLEDGMENT

This work is funded in part by the National Natural Science Foundation of China Project (No. 62372078, 62376049) and the Dalian Key Field Innovation Team Support Plan (Grant: 2020RT07).

REFERENCES

[1] P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," 2018.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," 2017.

[3] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic-and visual-lstm with channel-aware temporal convolution network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1424–1437, 2021.

[4] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," 2019.

[5] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," 2020.

[6] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[7] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Advances in neural information processing systems*, vol. 25, 2012.

[8] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," 2017.

[9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," 2016.

[10] G. Hu, Y. Hua, Y. Yuan, Z. Zhang, Z. Lu, S. S. Mukherjee, T. M. Hospedales, N. M. Robertson, and Y. Yang, "Attribute-enhanced face recognition with neural tensor fusion networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3744–3753.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[12] S. Mai, H. Hu, and 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*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 481–492. [Online]. Available: https://aclanthology.org/P19-1046

[13] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic- and visual-lstm with channel-aware temporal convolution network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1424–1437, 2021.

[14] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Transactions on Multimedia*, vol. 25, p. 4121–4134, 2023. [Online]. Available: http://dx.doi.org/10.1109/TMM.2022.3171679

[15] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L. philippe Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," 2021.

[16] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," 2021.

[17] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 169–176.

[18] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," 2018.

[19] S. Mai, H. Hu, and 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.

[20] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 122–137, 2019.

[21] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," 2020.

[22] S. Mai, S. Xing, J. He, Y. Zeng, and H. Hu, "Analyzing unaligned multimodal sequence via graph convolution and graph pooling fusion," 2021.

[23] A. B. Zadeh, P. 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[24] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," 2021.

[25] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," 2018.

[26] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," 2020.

[27] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "Cubemlp: An mlp-based model for multimodal sentiment analysis and depression estimation," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. ACM, Oct. 2022. [Online]. Available: http://dx.doi.org/10.1145/3503161.3548025

[28] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[29] ——, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[30] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[31] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," 2018.

[32] C. Lee and M. van der Schaar, "A variational information bottleneck approach to multi-omics data integration," 2021.

[33] Q. Wang, C. Boudreau, Q. Luo, P.-N. Tan, and J. Zhou, "Deep multi-view information bottleneck," in *Proceedings of the 2019 SIAM International Conference on Data Mining*. SIAM, 2019, pp. 37–45.

[34] R. Wang, X. He, R. Yu, W. Qiu, B. An, and Z. Rabinovich, "Learning efficient multi-agent communication: An information bottleneck approach," 2020.

[35] A. Goyal, R. Islam, D. Strouse, Z. Ahmed, M. Botvinick, H. Larochelle, Y. Bengio, and S. Levine, "Infobot: Transfer and exploration via the information bottleneck," 2019.

[36] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, "Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow," *arXiv preprint arXiv:1810.00821*, 2018.

[37] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[38] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," 2019.

[39] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 3718–3727.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014.

[43] A. Shapiro, "Monte carlo sampling methods," *Handbooks in operations research and management science*, vol. 10, pp. 353–425, 2003.

[44] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[45] A. B. Zadeh, P. 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[46] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

[47] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8992–8999.

[48] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," 2020.

[49] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[50] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.