Adaptive Language-Interacted Hyper-Modality Representation for Multimodal Sentiment Analysis

Lei Pan, WenLong Liu*
College of Computer, Zhongyuan University of Technology, Zhengzhou, Henan 450007, China

***Abstract*—In an attempt to mitigate the problem of neglecting unimodal information and incorporating emotionally unrelated data during the fusion process of multimodal representation, this study presents an adaptive language interaction representation (Adaptive Language-interacted Representation, ALR) model in this study. Initially, the unimodal representation module is utilized to obtain a minimal but adequate representation of the unimodal information. Subsequently, we acknowledge that video and audio modalities may contain sentiment data that is not relevant. To address this issue, hyper-modality representation is constructed to mute the impact of irrelevant sentimental information. This is achieved through interaction among text, video and audio features. Finally, the hyper-modality representation is integrated through multimodal fusion module, harnessing more efficient multimodal sentiment analysis. On the datasets CMU-MOSEI, MELD and IEMOCAP, the model outperforms the major of existing sentiment analysis models.**

Keywords*—Multimodal; multimodal fusion; sentiment analysis; adaptive language-interacted

I. Introduction

In recent years, the realm of multimodal sentiment analysis has gained considerable momentum within sentiment computing. Propelled by advancements in multimodal machine learning and dialogue systems, this area has become integral for equipping machines to perceive, recognize, and comprehend human behaviours and intentions [1] [2]. Beyond spoken words, individuals express opinions and emotions through various modalities, in which facial expressions and vocal cues play vital roles in both human-to-human and machine-to-machine communication. Exclusively relying on unimodal data for sentiment analysis is frequently inadequate for aptly capturing the genuine emotions expressed by individuals, thereby leading to potential misinterpretations. Multimodal sentiment analysis augments the amalgamation of information across various modalities and alleviates inherent ambiguities within individual modalities, thus yielding more precise and reliable model outcomes.

The foremost challenges in the field of multimodal sentiment analysis originate from the representation of unimodal data and the assimilation of cross-modal information. Previous research typically employed pre-trained models to elucidate features from individual modes and contrived sophisticated fusion techniques to assimilate multimodal embeddings, such as tensor fusion and Transformer-based fusion [3] [4]. Although these approaches prove to be effective, they are overly complex, and the resulting high-dimensional multimodal embeddings have a tendency for redundancy, thereby escalating the risk of overfitting. In an ideal scenario, multimodal embeddings should encapsulate the optimal amount of pertinent information indispensable for accurate forecasting while shedding extraneous data. In this research work, we posit that the multimodal embeddings, yielded by complex fusion networks, might encompass redundancies that outshine the crucial discriminative unimodal information. For instance, Zadeh et al. [3] utilized an outer product to generate a high-order multimodal tensor, resulting in a redundant representation that could potentially eclipse precious unimodal information during the forecasting process. Moreover, multiple research instances and corresponding ablation experiments have established the differential contributions of various modalities to emotion recognition, with linguistic aspects often assuming a paramount role [5]. We further note the presence of ambiguities and contradictions within information derived from differing modalities, specifically non-dominant ones such as illumination and action postures in videos, or background noise in audio recordings. These contentious pieces of data can significantly undermine the proficiency of multimodal sentiment analysis.

To resolve these identified problems, an avant-garde ALR model is put forward in this paper. The model encompasses a unimodal representation module, calibrated to elicit individual modalities, thereby slenderising each modality by eliminating disruptive information and retaining modality-specific data. Conversely, a textual interaction module utilises prevailing linguistic characteristics to converse with various video and audio modalities, thereby deriving the final modality data. This data, which comprises minimal emotionally inconsequential elements, augments the recognition of essential emotional attributes, thereby bolstering the sentiment analysis efficacy of the model. The primary contributions of this work are articulated as follows:

1) Employing the principle of mutual information, the posited methodology models the data within the unimodal state, effectively filtering out noise while safeguarding distinctive information. This refinement markedly elevates the model's proficiency in emotion recognition.

2) We have carved an effective feature representation that leverages linguistic attributes for an interplay with video and audio characteristics. This facilitates the creation of a comprehensive multimodal representation, mitigating modality discrepancies. The resultant advantage is a superior model capacity in recognising critical emotional traits.

3) Rigorous comparative and ablation tests executed on three extensively utilized multimodal sentiment analysis benchmark datasets—specifically CMU-MOSEI, MELD, and IEMOCAP—unequivocally demonstrate ALR's superior

performance over prior techniques in the most evaluation criteria.

The rest of this paper is as follows, Section II review previous studies. Section III discusses the methodology. Section IV presents experimental setup. Section V describes the results of the experiment and discusses. Finally, conclusion presents in Section VI.

II. Related Works

In this segment, we succinctly examine precedents from two vantage points: multimodal sentiment analysis and Transformers.

A. Multimodal Sentiment Analysis

Multimodal sentiment analysis is rooted in the burgeoning interdisciplinary field that intersects natural language processing, computer vision, and speech recognition. Prior techniques for multimodal sentiment analysis fall typically into three broad categories: ones centered on representation learning, those concentrated on multimodal fusion, and methods focused on pre-trained models.

As for representation learning-centered methods, Hazarika et al. [5] and Yang et al. [6] treated multimodal representation learning as a domain adaptive task and attained leading-edge results across a range of datasets. They utilized metric and adversarial learning to harness modality-invariant and modality-specific representations for multimodal fusion. Proposed by Pham et al. [7], the Multimodal Cyclic Translation Network (MCTN) learns robust conjoint multimodal representations by implementing cross-modal translation. Guo et al. [8] amalgamated both linguistic and non-linguistic behavioural data to secure enhanced linguistic representations. Moreover, Wang et al. [9] put forward recursive attention change embedding networks to induce multimodal shifts. Nevertheless, these approaches fall short in sufficiently addressing the presence of superfluous information unrelated to emotion within video and audio modalities, thereby limiting the performance of model.

Regarding multimodal fusion-focused methods, Sun et al. [10] brought forth a two-stage multimodal fusion blueprint titled TIMF, which deftly meshes both initial and subsequent fusion mechanisms for sentiment analysis undertakings. On a different note, Tsai et al. [4] brought forward the Multimodal Transformer, an approach designed to align sequences and to harness long-range interdependencies amongst cross-modal elements. Liang et al. [11] advanced the Recursive Multi-stage Fusion Network (RMFN), a framework that dissects the multimodal fusion issue into several iterative stages. Every phase pays close attention to a unique subset of multimodal attributes, paving the way for efficient intermodal fusion. Nevertheless, such methods centre predominantly on blending data from singular modalities, leading to the possible inclusion of emotionally non-pertinent data, thus bringing about less than ideal results.

In the area of pre-trained model-focused techniques, Ando et al. [12] advanced a sequential cross-modal model, dubbed UEGD. Here, video, audio, and text are duly encoded utilizing tools such as the CLIP Vision Transformer [13], WavLM [14], and BERT [15]. Afterwards, the conjoint representation of the information from these trio of modalities is achieved via gating units. Aziz et al. [16] put forward a multimodal Transformer, dubbed as MMTF-DES. This technique acquires the contextual representation of video and language by collaboratively fine-tuning both the video-language Transformer and the video-enhanced language Transformer. It then employs an early fusion approach to secure the feature representation of the image-text pairing. The objective of the above methods hinges on extracting modal features via the utilization of pre-trained models, followed by attaining inter-modal fusion through a simplistic fusion strategy. Nonetheless, these methodologies overlook the factor of inter-modal variability, and non-verbal modalities may encompass disruptive noise, consequently impeding the performance of the model.

B. Transformer

The Transformer, introduced by Vaswani et al. [17], is an advanced machine translation model that leverages attention mechanisms. Depicting a sequence-to-sequence model devoid of any recurrent structures, it exhibits outstanding modelling capabilities across multiple tasks including but not limited to natural language processing, computer vision, and language processing [18]. This technique has been proficiently employed in multimodal sentiment analysis for the purpose of feature extraction, representation learning, and multimodal fusion [19].

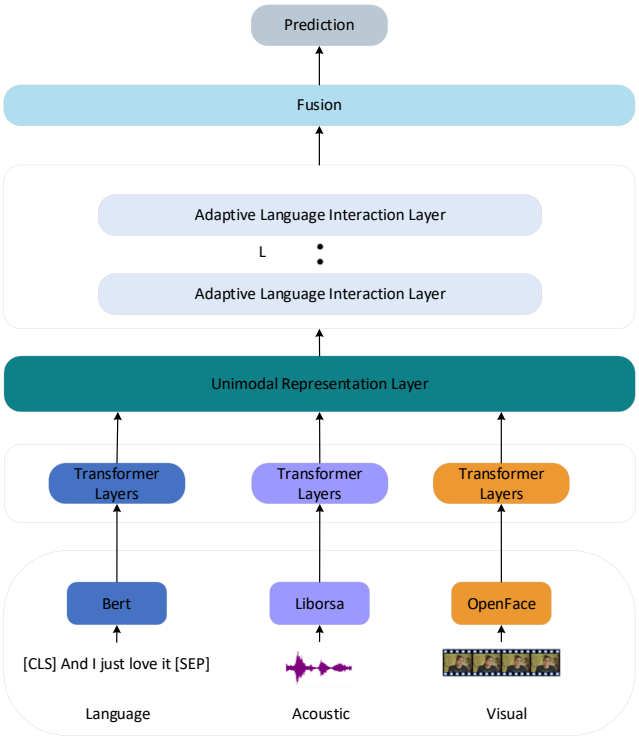

Fig. 1. ALR model structure framework.

III. Methodology

A. Overview of the Model

In this study, we present an adaptive language interaction representation (ALR) model for multimodal sentiment analysis is in Fig. 1. As shown, ALR first extracts uniform modal features from input. Then, model embedding is performed on the modal features. The Unimodal Representation (UR) module is used to learn the minimum adequate representation of the unimodal modality and eliminate the redundant information within the

modality. The Adaptive Language Interaction (ALI) module is used to learn adaptive hyper-modality representation dominated by linguistic features at different scales. Finally, we apply a Modal Fusion module to synthesize the hyper-modality features with language features, thus obtaining a language interaction representation model for multimodal sentiment analysis.

B. Multimodal Input

When dealing with multimodal inputs, the approach presented in this paper involves the extraction of features from text, audio, and video through BERT, Librosa [20], and OpenFace [21]. These features are represented as $U_m \in R^{T_m \times d_m}$ where $m \in \{l, a, v\}$ with T_m representing the sequence length and d_m indicating the feature dimensions. It's important to note that in real-world applications, different modalities within the dataset may have varying sequence lengths and feature dimensions.

C. Modality Embedding

In the modality embedding, we introduce Transformer layer. These layers are designed to capture temporal features from each modality, as depicted in Eq. (1).

$$x^*_m = Transformer(x_m) \tag{1}$$

Where, x_m is the initial feature sequence of three modalities, x^*_m is the feature sequence after encoding.

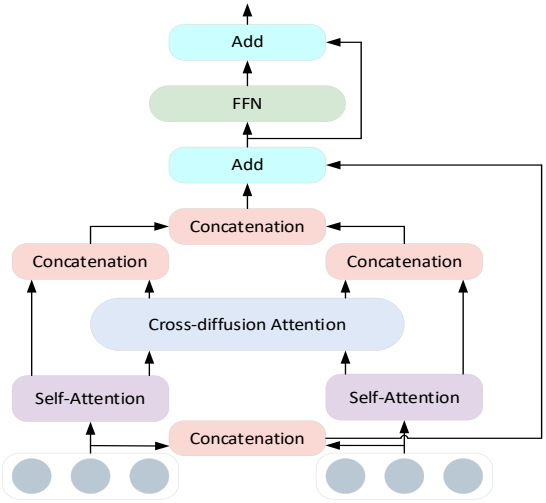

Fig. 2. Adaptive language interaction structure diagram.

D. Unimodal Representation

In the realm of unimodal representation, the concept of Information Bottleneck (IB) is introduced. The IB framework seeks to obtain improved representations within the constraints of complexity. It aims to ensure that the representations are both discriminative and free from redundant information. The IB approach defines the quality of a representation based on a fundamental trade-off between conciseness and predictive power. It utilizes Mutual Information (MI) as a basis and strives to maximize the MI between the coded representations and the corresponding labels, while minimizing the MI between the coded representations and the input data. By striking a balance between these two objectives, the IB framework aims to derive representations that are both informative and efficient.

MI is a concept used to quantify the interdependence between two random variables. It measures the amount of information that one variable provides about the other. If the values of two variables are completely independent, their mutual information is zero. Conversely, if the values of the variables are highly correlated, the mutual information is maximized. Formally, given two random variables x and y, they have a joint distribution $p(x, y)$ and marginal distributions $p(x)$ and $p(y)$. Their MI is defined as the Kullback-Leibler (KL) divergence between the joint distribution and the marginal product, as depicted in Eq. (2).

$$\begin{aligned} I(x;y) = I(y;x) &= KL\big(p(x,y) \,\|\, p(x)p(y)\big) \\ &= \int dxdy p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{2}$$

The goal of IB is to use the input x to learn the compressed coded representation z, where z is maximally discriminative with respect to the target variable y (i.e. $I(y;z)$ is maximised). Clearly, the most informative representation can be obtained by the same mapping (i.e., $x = z$), but this mapping contains noise, which is redundant information for prediction. Therefore, a MI constraint is added between z and x, so the goal of the information bottleneck becomes:

$$\max I(y;z) \tag{3}$$

$$\min I(x;z) \tag{4}$$

The first constraint in the Information Bottleneck (IB) framework aims to maximize the prediction of the target variable. On the other hand, the second constraint aims to minimize the inclusion of information from the target variable. In essence, the goal of IB is to learn a representation that contains only the essential information that is discriminative for accurate prediction. The objective function of the Information Bottleneck can be expressed as follows:

$$F_{IB} = I(y;z) - \beta I(x;z) \tag{5}$$

The weight of the minimum information constraint, a scalar denoted as β, plays a crucial role in determining its influence during the optimization process, we set the value of β to 1 default. The minimum adequate representation of each modality is obtained through the unimodal representation layer, which denoted as $\overline{\overline{x_m}}$, is used as the initial input to the adaptive language interaction layer.

E. Adaptive Language Interaction

In this study, we introduce an adaptive language interaction layer, whose overall structure is shown in Fig. 2. The text modality is interacted with audio and video modalities respectively to obtain a feature representation that suppresses emotionally irrelevant information.

We represent the feature vectors of Modal-r and Modal-d as $X_r \in R^{n\times d}$ and $X_d \in R^{n\times d}$, where Modal-r and Modal-d denote the two different modal of the input adaptive language interaction module. Here, n represent the length of modal sequence, and d represent the features of dimension. To obtain the dependency of tokens within each modal, Self-Attention is used for each modal. First, the correlation between different tokens of Modal-r is calculated:

$$\alpha_r = soft\max(\frac{Q_r K_r^T}{\sqrt{d}}) \tag{6}$$

where Q_r and K_r are obtained by making linear variations of X_r, $\sqrt{d}$ denote the scaling factor. The context-aware representation of Modal-r is obtained through the message passing mechanism based on α_r, as follows:

$$\overline{X_r} = \alpha_r V_r \tag{7}$$

where V_r are obtained by making linear change X_r. It is also possible to get α_d and $\overline{X_d}$ for Modal-d.

Interaction of Modal-r and Model-d by Cross diffusion Attention (CDA) [22], as follows:

$$\overline{X}_{d\to r} = CDA(\overline{X}_d, \overline{X}_r) \tag{8}$$

$$\overline{X}_{r\to d} = CDA(\overline{X}_r, \overline{X}_d) \tag{9}$$

We obtain H_d and H_r by concatenating $\overline{X}_d$ with $\overline{X}_{d\to r}$ and $\overline{X}_r$ with $\overline{X}_{r\to d}$, as follows:

$$H_d = F_d(\overline{X}_d \,\|\, \overline{X}_{d\to r}) \tag{10}$$

$$H_r = F_r(\overline{X}_r \,\|\, \overline{X}_{d\to r}) \tag{11}$$

where $\|$ represents the splicing operation in the channel dimension, $F_d(\square)$ and $F_r(\square)$ represent two convolutional layers with different parameters.

Finally, H_d and H_r are aggregated together and then the hyper-modality representation P is obtained through forward feedback network.

$$H = g(H_r \,\|\, H_d) + h(X_r \,\|\, X_d) \tag{12}$$

$$P = FFN(H) + H \tag{13}$$

where $g(\square)$ and $h(\square)$ represent two convolutional layers with different parameters and $FFN(\square)$ represents a single fully connected layer with nonlinear activation function.

F. Multimodal Fusion and Output

We can obtain hyper-modality representation of video and audio through adaptive language interaction module. Subsequently, we fused the video hyper-modality representation P_v, the audio hyper-modality representation P_a and the textual modality representation $\overline{\overline{x_l}}$ through modal fusion to get the final vector U for sentiment analysis, as follows:

$$U = Fusion(P_v, P_a, \overline{\overline{x_l}}) \tag{14}$$

For CMU-MOSEI, a single fully connected layer is used for linear transformation to obtain the final sentiment value prediction. The model is optimized using the mean absolute error as the loss function, as follows:

$$y^* = FFN(U) \tag{15}$$

$$Loss = \frac{1}{N}\sum_{i=1}^{N} | y_i - y_i^* | \tag{16}$$

where N is the total number of samples, i is the sample serial number, y_i is the true sentiment value and y_i^* is the predicted sentiment value.

For MELD and IEMOCAP, a single fully connected layer is used for linear transformation to obtain the final sentiment categories. The model is optimized using the cross entropy as the loss function, as follows:

$$y^* = FFN(U) \tag{17}$$

$$Loss = \sum_{j=1}^{N} y_j \log(y^*_j) \tag{18}$$

where N is the total number of samples, j is the sample serial number, y_j is the true sentiment category and y_j^* is the predicted sentiment category.

IV. Experimental Setup

A. Datasets and Evaluation Metrics

1) Datasets: We conducted extensive experiments on three popular datasets, the details of which are shown in Table I:

CMU-MOSEI [23] is a large multimodal sentiment analysis dataset containing a total of 22,856 YouTube movie review clips. Each discourse is scored into two levels, sentiment scores ranging from [-3, 3].

MELD [24] comprises 13,707 video dialogue clips with labels following Ekman's six universal emotions containing joy, sadness, fear, anger, surprise, and disgust.

IEMOCAP [25] consists of 7,532 samples. Following previous works selected from six emotions including joy, sadness, anger, neutral, excited, and frustrated.

TABLE I. Details of each dataset

Dataset	Train	Valid	Test	All
CMU-MOSEI	16326	1871	4659	22856
MELD	9989	1108	2610	13707
IEMOCAP	5354	528	1650	7532

2) Evaluation metrics: For the CMU-MOSEI dataset, we adhere to established methodologies by employing mean absolute error (MAE), which represents the average absolute difference between predicted and actual values. We also use Pearson correlation (Corr) to gauge the degree of prediction bias, seven-class classification accuracy (Acc-7) to measure the proportion of predictions that correctly fall within the same interval of seven ranges between -3 and +3 as the actual values, and binary classification accuracy (Acc-2) along with the F1 score for positive/negative classification results. For the MELD and IEMOCAP datasets, we utilize accuracy (Acc) and weighted F1 (WF1) for evaluation. The WF1 is a multi-category assessment metric that accounts for category imbalances by weighting the average F1-score for each category.

B. Experimental Details

We develop our model using PyTorch on RTX4060Ti with CUDA 12.1 and torch 2.10. Following a randomized search for optimal hyperparameters, we selected the test outcomes corresponding to the most favourable configuration as our reported results. The specific model parameters are detailed in Table II, and we used a random seed value of 1024 for reproducibility. To mitigate the risk of overfitting during training, we implemented an early stopping technique. Furthermore, we used the Adam optimizer to facilitate the learning process.

TABLE II. DETAILS OF EXPERIMENTAL PARAMETERS

Parameter	Value
Epoch	50
Learning Rate	1e-5
Dropout	0.5
Batch Size	64
Optimizer	Adam

C. Baseline

To comprehensively validate the performance of our ALR, we make a fair comparison with the several advanced and state-of art methods, and the following benchmark models are involved in this study:

- TFN [3]. The interactions among unimodal, bimodal, and trimodal elements are achieved through the computation of outer products within the trimodal tensor.
- LMF [26]. The approach being proposed utilises a low-rank tensor decomposition method designed for effective multimodal fusion, which significantly decreases the computational complexity inherent in the integration process.
- MFN [27]. The method being proposed harnesses the potential of Long Short-Term Memory (LSTM) networks, enabling the encoding of temporal interactions contained in multimodal sequences. Following this, the Dynamic Multimodal Attention Network (DMAN) is engaged to pinpoint and incorporate cross-modal connections. Lastly, the LSTM structure is again applied to capture and refresh the information of the advanced multimodal sequence.
- MM-DFN [28]. The approach makes use of an adapted graph convolutional neural network for the amalgamation of multimodal contextual characteristics. This results in a decrease in redundancy and an enhancement of inter-modal complementarity, accomplished by the capture of contextual information across varied semantic spaces.
- RAVEN [10]. The method is designed to learn non-linear combinations of video and audio embeddings through an attention mechanism, leading to the calculation of non-verbal offset vectors for temporal modelling. Subsequently, these offset vectors come into play to fine-tune the representations of words.
- MULT [4]. The technique harnesses a directional cross-modal attention mechanism, promoting interplay across multimodal sequences at varying temporal junctures. This, in turn, creates an avenue for potential adaptability from one modality to another.
- MFM [29]. The method presents a novel method for multimodal feature depiction, achieving this by segregating each information mode into shared discriminators and distinct generators.
- IMR [30]. The method proactively adjusts the weightage between the input modality and the output characterisation, implementing individualised tweaks for every given input sample.
- QMF [31]. The method unveils a novel structure, which borrows insights from quantum theory, with the intent to address the constraints of neural networks by applying a technique rooted in interaction and correlation.
- MISA [5]. The approach breaks down modal representations into modal invariant and modal specific depictions, employing a metric-oriented strategy to maintain consistency and variability amongst them.
- DialogueGCN [32]. The method leverages the power of graphical convolutional neural networks to tackle the issue of context propagation, using dependency modelling to bridge the conversational gap between the dialogue parties.
- COSMIC [33]. The approach harnesses commonsense information garnered from the dialogue, including the speaker's reactions, emotional states, and intentions.
- MAG-BERT [34]. The proposed method integrates BERT and XLNet with the Multimodal Attention Gateway (MAG), enabling the assimilation of multimodal non-verbal data during the fine-tuning phase.
- UniMSE [35]. The strategy incorporates a pre-trained Modal Fusion layer (PMF) into the Transformer tier of T5, fusing textual features at diverse degrees with audio and video data, to access a rich array of information. Moreover, cross-modal comparison learning is carried

out to diminish the intra-modal variation and simultaneously amplify the inter-modal differential.

- HCT-MG [36]. The strategy adeptly discerns the primary modality and coordinates hierarchical exchanges between the primary and secondary modalities, thereby proficiently reducing redundancy amongst the modalities.

V. Results and Discussion

A. Model Comparison Experiment

Table III and Table IV showcase the comparative results of both the precedent benchmark model as mentioned in the preceding subsection and the model proposed in this study, using equivalent evaluation metrics, on the CMU-MOSEI, MELD and IEMOCAP datasets. The result in Table III and Table IV are based on MMMU-BA [37] as fusion method.

Table III and Table IV list the comparison results of our proposed method and state-of-the-art methods on CMU-MOSEI, MELD and IEMOCAP, respectively. As shown in the Table III, the proposed ALR achieve competitive performance in most evaluation metrics. On the task of more difficult sentiment classification (Acc-7), our model achieves remarkable improvements. For example, on the CMU-MOSEI dataset, ALR achieved a relative improvement of 1.5% compared to the result obtained by MISA. It demonstrates that the elimination of noise within a single mode and redundant information in cross-modal interactions is essential for multimodal sentiment analysis.

TABLE III. Comparison with Baselines on CMU-MOSEI

Method	CMU-MOSEI				
	Acc-7	Acc-2	F1	MAE	Corr
TFN	49.80	79.40	79.70	0.610	0.671
LMF	50.00	80.60	81.00	0.608	0.677
MFN	49.10	79.60	80.60	0.618	0.670
RAVEN	50.20	79.00	79.40	0.605	0.680
MULT	48.20	80.20	80.50	0.638	0.659
MFM	51.30	84.40	84.30	0.568	0.703
IMR	48.70	80.60	81.00	-	-
QMF	47.90	80.70	79.80	0.640	0.658
MISA	52.20	**85.50**	85.30	0.555	0.756
MAG-BERT	51.90	85.00	85.00	0.602	0.778
UniMSE	48.68	-	-	0.691	**0.809**
HCT-MG	50,60	81.60	81.90	0.593	0.691
ALR	**53.70**	84.70	**85.70**	**0.541**	0.785

Moreover, it is worth noting that the scenarios in MELD and IEMOCAP are more complex the CMU-MOSEI. Therefore, it is more challenging to model the multimodal data. However, as shown in the Table IV, ALR achieve state-of-the-art performance in all metrics compared to the sub-optimal approach. For example, compared to UniMSE, it achieved relative improvement with 2.01% on Acc and 2.82% on the corresponding WF1 on MELD. Achieving such superior performance on MELD and IEMOCAP with more complex scenarios demonstrates ALR's ability to extract effective sentiment information from various scenarios.

TABLE IV. Comparison with Baselines on MELD and IEMOCAP

Method	MELD		IEMOCAP	
	Acc	WF1	Acc	WF1
TFN	60.70	57.74	55.02	55.13
LMF	60.70	57.74	56.50	56.49
MM-DFN	62.49	59.46	68.21	68.18
MFM	60.08	57.80	61.24	61.60
DialogueGCN	59.46	58.10	65.25	64.18
COSMIC	-	65.21	-	65.28
UniMSE	65.09	65.51	70.56	70.66
ALR	**67.10**	**68.33**	**72.10**	**71.80**

B. Analysis of Ablation Experiments

1) Effects of different modalities: To better understand the influence of each modality in the proposed ALR, Table V reports the ablation results of the subtraction of each modality to the ALR on the CMU-MOSEI dataset, respectively. We can find that removing visual and acoustic modalities or one of them all leads to performance degradation, which indicates that the non-verbal signals are necessary for solving multimodal sentiment analysis, and demonstrates the complementarity among text, acoustic, and visual.

TABLE V. Effects of Different Modalities

Method	MAE	Acc-2	Acc-7	F1	Corr
-w/o A	0.579	82.70	49.07	82.11	0.719
-w/o V	0.585	82.50	48.88	81.38	0.712
-w/o A, V	0.601	81.80	45.96	79.62	0.691
ALR	**0.541**	**84.70**	**53.70**	**85.70**	**0.785**

2) Effects of different components: To verify the effectiveness of each component of the proposed ALR, in Table VI, we present the ablation results of the subtraction of each component on the CMU-MOSEI dataset, respectively. ALR w/o UR, ALR w/o ALI models respectively remove the unimodal representation module, the adaptive language interaction module. We can find that deactivating the Unimodal Representation (UR) layer greatly decreases the performance, demonstrating the unimodal representation learning strategy is effective. Moreover, after the removal of the Adaptive Language Interaction (ALI) layer, the performance drops again, also supporting that the ALI layer can effectively improve the ALR's ability to interact with emotional information in each modality.

TABLE VI. Effects of Different Components

Method	MAE	Acc-2	Acc-7	F1	Corr
ALR w/o UR	0.577	81.49	50.88	81.38	0.712
ALR w/o ALI	0.585	82.21	51.07	82.11	0.719
ALR	**0.541**	**84.70**	**53.70**	**85.70**	**0.785**

3) Effects of different fusion methods: To substantiate the prowess of our proposed approach, we have amalgamated ALR with diverse fusion methods. The empirical outcomes are delineated in Table VII. The findings illustrate that the ALR model, as proposed herein, is amenable to a wide array of fusion techniques and delivers a superior depiction of modal attributes. As can be inferred from the tabulated data, the model exhibits enhanced performance when a sophisticated fusion mechanism is employed. This suggests that the ALR model has the capability to filter out noise within the modal representation and capture a sufficient encapsulation of the modal information.

TABLE VII. EFFECTS OF DIFFERENT FUSION METHODS

Method	MAE	Acc-2	Acc-7	F1	Corr
Concatenation	0.557	83.31	52.56	83.44	0.771
Addition	0.558	82.39	52.65	82.42	0.750
Multiplication	0.558	84.00	52.52	83.70	0.773
MMM-BA	**0.541**	**84.70**	**53.70**	**85.70**	**0.785**

C. Parameter Analysis

In this study, we experimented and analysed two important parameters for five evaluation metrics: the MAE, Acc-2, Acc-7, F1 and Corr values. One of the parameters examines the effect of the number of ALI layers on the model performance. The other parameter examines the effect of modal vector dimension on model performance.

1) Effects of different number of alI layer: In Table VIII, we experimented with different layers of ALI on CMU-MOSEI dataset. Probing the empirical data presented within the table, it is discernible that the model attains its peak performance indices when the count of ALI layers is contained to six layers. This observation suggests that an insufficient number of ALI layers results in a partial interaction between text features and audio-visual attributes. Conversely, an excessive number of layers induces an overbearing influence of the text features on audio-visual characteristics, thereby disregarding the discriminative information inherently present within audio-visual features.

TABLE VIII. EFFECTS OF DIFFERENT NUMBER OF ALL LAYER

ALI Layer	MAE	Acc-2	Acc-7	F1	Corr
3	0.549	84.42	53.17	84.23	0.772
6	**0.541**	**84.70**	**53.70**	**85.70**	**0.785**
9	0.555	82.67	53.00	83.82	0.761

2) Effects of different vector dimensions: In Table IX, we experimented with different vector dimensions on CMU-MOSEI dataset. The dimensionality of feature vectors directly impacts the magnitude and expressivity of the model. Employing higher-dimensional feature vectors proffers a wealth of data, thereby augmenting the model's propensity to discern complex inter-relationships; however, this necessitates more data for effectual training. On the other hand, utilizing lower-dimensional feature vectors has the potential to precipitate model underfitting and could fail to encapsulate intricate data patterns. As discerned from the experimental data tabulated, the model manifests optimum performance when the dimension of the feature vector is set at 256.

TABLE IX. EFFECTS OF DIFFERENT VECTOR DIMENSION

Vector Dimension	MAE	Acc-2	Acc-7	F1	Corr
128	0.556	84.30	51.62	84.28	0.761
256	**0.541**	**84.70**	**53.70**	**85.70**	**0.785**
512	0.5505	83.94	53.06	83.89	76.23

VI. CONCLUSION

This paper proposes an innovative Adaptive Language-interacted Representation (ALR) model earmarked, for multimodal sentiment analysis tasks. The essence of this model pivots on multimodal feature representations. Specifically, it constructs unimodal representations that leverages the concept of information bottlenecks to secure the most compressed yet efficient representation of unimodal data. The model further integrates employ text modality with video and audio modality, yielding a refined abstraction known as hyper-modality representations, which filter out emotionally insignificant features. The modal representation is ascertained via a combination of unimodal representation and textual interplay and is deemed a sufficient representation of the modal data. The proposed model delivers promising, if not superior outcomes, across various metrics. This underscores the significance of generating a minimal, yet effective, amalgamation of feature representations, a vital aspect enhancing sentiment prediction efficacy.

In the field of multimodal sentiment analysis, an inclusive amalgamation of multimodal data holds significant importance. However, the disparate distributions of sentiment data across diverse modalities significant challenge to achieving the optimal integration of modal information. As a result, future work seeks to establish a multimodal dynamic fusion network, with a purpose to dynamically interlink different modal information. It is expected that this approach will not only facilitate the comprehensive fusion of data across various modalities but also enrich the representation of the resulting fused features.

REFERENCES

[1] P. P. Liang, A. Zadeh & L. P. Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions[J]. arXiv preprint arXiv:2209.03430, 2022.

[2] H. L. Zhang, H. Xu & T. E. Lin. Deep open intent classification with adaptive decision boundary[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 35(16): pp. 14374-14382, 2021.

[3] A. Zadeh, M. Chen, S. Poria, E. Cambria & L. P. Morency. Tensor fusion network for multimodal sentiment analysis[J]. arXiv preprint arXiv:1707.07250, 2017.

[4] Y. H. H. Tsai, S. Bai & P. P. Liang, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, p. 6558, 2019.

[5] D. Hazarika, R. Zimmermann & S. Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis[C]//Proceedings of the 28th ACM international conference on multimedia, pp. 1122-1131, 2020.

[6] D. Yang, S. Huang, H. Kuang, Y. T. Du & L. H. Z. Disentangled representation learning for multimodal emotion recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia, pp. 1642-1651, 2022.

[7] H. Pham, P. P. Liang, T. Manzini, L. P. Morency & B. Poczos. Found in translation: Learning robust joint representations by cyclic translations between modalities[C]//Proceedings of the AAAI conference on artificial intelligence, pp. 6892-6899, 2019.

[8] J. Guo, J. Tang, W. Dai, Y. Ding & W. Kong. Dynamically adjust word representations using unaligned multimodal information[C]//Proceedings of the 30th ACM International Conference on Multimedia, pp. 3394-3402, 2022.

[9] Y. Wang, Y. Shen & Liu Z, et al. Words can shift: Dynamically adjusting word representations using nonverbal behaviors[C]//Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7216-7223, 2019.

[10] J. Sun, H. Yin & Y. Tian, et al. Two-level multimodal fusion for sentiment analysis in public security[J]. Security and Communication Networks, pp. 1-10, 2021.

[11] P. P. Liang, Z. Liu, A. Zadeh & L. P. Morency. Multimodal language analysis with recurrent multistage fusion[J]. arXiv preprint arXiv:1808.03920, 2018.

[12] A. Ando, R. Masumura & A. Takashima, et al. On the use of modality-specific large-scale pre-trained encoders for multimodal sentiment analysis[C]//2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 739-746, 2023.

[13] A. Radford, J. W. Kim & C. Hallacy, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, pp. 8748-8763 2021.

[14] S. Chen, C. Wang & Z. Chen, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, pp. 1505-1518, 2022.

[15] J. Devlin, M. W. Chang, K. Lee & K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[16] A. Aziz, N. K. Chowdhury, M. A. Kabir, A. N. Chy & M. J. Siddique. MMTF-DES: A Fusion of Multimodal Transformer Models for Desire, Emotion, and Sentiment Analysis of Social Media Data[J]. arXiv preprint arXiv:2310.14143, 2023.

[17] A. Vaswani, N. Shazeer & N. Parmar, et al. Attention is all you need[J]. Advances in neural information processing systems, p. 30, 2017.

[18] Y. Liu, W. Wang & C. Feng, et al. Expression snippet transformer for robust video-based facial expression recognition[J]. Pattern Recognition, p. 109368, 2023.

[19] Y. Liu, H. Zhang & Y. Zhan, et al. Noise-resistant multimodal transformer for emotion recognition[J]. arXiv preprint arXiv:2305.02814, 2023.

[20] B. McFee, C. Raffel & D. Liang, et al. librosa: Audio and music signal analysis in python[C]//SciPy, pp. 18-24, 2015.

[21] T. Baltrusaitis, A. Zadeh, Y. C. Lim & L. -P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, pp. 59-66, 2018.

[22] X. Wang, X. Wang, B. Jiang, J. Tang & B. Luo. MutualFormer: Multi-Modality Representation Learning via Cross-Diffusion Attention[J]. arXiv preprint arXiv:2112.01177, 2021.

[23] A. A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria & L. P. Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236-2246, 2018.

[24] S. Poria, D. Hazarika & N. Majumder, et al. Meld: A multimodal multi-party dataset for emotion recognition in conversations[J]. arXiv preprint arXiv:1810.02508, 2018.

[25] C. Busso, M. Bulut & C. C. Lee, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language resources and evaluation, pp. 335-359, 2008.

[26] Z. Liu, Y. Shen & V. B. Lakshminarasimhan, et al. Efficient low-rank multimodal fusion with modality-specific factors[J]. arXiv preprint arXiv:1806.00064, 2018.

[27] A. Zadeh, P. P. Liang & N. Mazumder, et al. Memory fusion network for multi-view sequential learning[C]//Proceedings of the AAAI conference on artificial intelligence, p. 32, 2018.

[28] D. Hu, X. Hou, L. Wei, L. Jiang & Y Mo. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7037-7041, 2022.

[29] Y. H. H. Tsai, P. P. Liang, A. Zadeh, L. P. Morency, R. Salakhutdinov. Learning factorized multimodal representations[J]. arXiv preprint arXiv:1806.06176, 2018.

[30] Y. H. H. Tsai, M. Ma, M. Yang, R. Salakhutdinov & L. P Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. NIH Public Access, p. 1823, 2020.

[31] Q. Li, D. Gkoumas, C. Lioma & M. Melucci. Quantum-inspired multimodal fusion for video sentiment analysis[J]. Information Fusion, pp. 58-71, 2021.

[32] D. Ghosal, N. Majumder, S. Poria, N. Chhaya & A. Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation[J]. arXiv preprint arXiv:1908.11540, 2019.

[33] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea & S. Poria. Cosmic: Commonsense knowledge for emotion identification in conversations[J]. arXiv preprint arXiv:2010.02795, 2020.

[34] W. Rahman, M. K. Hasan & S. Lee, et al. Integrating multimodal information in large pretrained transformers[C]//Proceedings of the conference. Association for Computational Linguistics. Meeting. NIH Public Access, p.2359, 2020.

[35] G. Hu, T. E. Lin & Y. Zhao, et al. Unimse: Towards unified multimodal sentiment analysis and emotion recognition[J]. arXiv preprint arXiv:2211.11256, 2022.

[36] Y. Wang, Y. Li & P. Bell, et al. Cross-attention is not enough: Incongruity-aware multimodal sentiment analysis and emotion recognition[J]. arXiv preprint arXiv:2305.13583, 2023.

[37] D. Ghosal, M. S. Akhtar & D. Chauhan, et al. Contextual inter-modal attention for multi-modal sentiment analysis[C]//proceedings of the 2018 conference on empirical methods in natural language processing, pp. 3454-3466, 2018.