UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition

Guimin Hu[†], Ting-En Lin, Yi Zhao[†][*], Guangming Lu[†], Yuchuan Wu, Yongbin Li[*]

[†]School of Computer Science and Technology, School of Science,
Harbin Institute of Technology (Shenzhen), China

{rice.hu.x,liyb821}@gmail.com
{zhao.yi,luguangm}@hit.edu.cn

Abstract

Multimodal sentiment analysis (MSA) and emotion recognition in conversation (ERC) are key research topics for computers to understand human behaviors. From a psychological perspective, emotions are the expression of affect or feelings during a short period, while sentiments are formed and held for a longer period. However, most existing works study sentiment and emotion separately and do not fully exploit the complementary knowledge behind the two. In this paper, we propose a multimodal sentiment knowledge-sharing framework (UniMSE) that unifies MSA and ERC tasks from features, labels, and models. We perform modality fusion at the syntactic and semantic levels and introduce contrastive learning between modalities and samples to better capture the difference and consistency between sentiments and emotions. Experiments on four public benchmark datasets, MOSI, MOSEI, MELD, and IEMO-CAP, demonstrate the effectiveness of the proposed method and achieve consistent improvements compared with state-of-the-art methods.

1 Introduction

With the rapid development of multimodal machine learning (Liang et al., 2022; Baltrušaitis et al., 2018) and dialog system (He et al., 2022a,b,c), Multimodal Sentiment Analysis (MSA) and Emotion Recognition in Conversations (ERC) have become the keys for machines to perceive, recognize, and understand human behaviors and intents (Zhang et al., 2021a,b; Hu et al., 2021a,b). Multimodal data provides not only verbal information, such as textual (spoken words) features but also non-verbal information, including acoustic (prosody, rhythm, pitch) and visual (facial attributes) features. These different modalities allow the machine to make decisions from different perspectives, thereby achieving more accurate predictions (Ngiam et al., 2011). The goal of MSA is to predict sentiment intensity
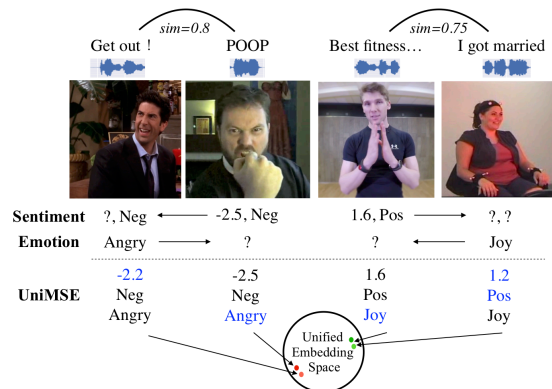
---
[*]Corresponding authors.

Figure 1: Illustration of sentiment and emotion sharing a unified embedding space. The bottom is a unified label after formalizing sentiment and emotion according to the similarity $\widehat{sim}$ between samples with the same sentiment polarity label.

or polarity, and ERC aims to predict predefined emotion categories. There are many research directions in MSA and ERC, such as multimodal fusion (Yang et al., 2021), modal alignment (Tsai et al., 2019a), context modeling (Mao et al., 2021) and external knowledge (Ghosal et al., 2020). However, most existing works treat MSA and ERC as separate tasks, ignoring the similarities and complementarities between sentiments and emotions.

On the one hand, from a psychological perspective, both sentiments and emotions are experiences that result from the combined influences of the biological, cognitive, and social (Stets, 2006), and could be expressed similarly. In Figure 1, we illustrate how sentiments and emotions are relevant in the verbal or non-verbal, and could be projected into a unified embedding space. On the other hand, emotions are reflections of the perceived change in the present within a short period (Batson et al., 1992), while sentiments are held and formed in longer periods (Murray and Morgan, 1945). In our preliminary study, we found that the video dura-

tion of MSA is almost twice of ERC [1], which is consistent with the above definitions. A variety of psychological literature (Davidson et al., 2009; Ben-Ze'ev, 2001; Shelly, 2004) explain the similarities and differences between sentiment and emotion. Munezero et al. (2014) also investigates the relevance and complementarity between the two and point out that analyzing sentiment and emotion together could better understand human behaviors.

Based on the above motivation, we propose a multimodal sentiment knowledge-sharing framework that **Uni**fied **MS**A and **E**RC (UniMSE) tasks. UniMSE reformulates MSA and ERC as a generative task to unify input, output, and task. We extract and unify audio and video features and formalize MSA and ERC labels into Universal Labels (UL) to unify sentiment and emotion.

Besides, previous works on multimodal fusion at multi-level textual features (Peters et al., 2018; Vaswani et al., 2017), like syntax and semantics, are lacking. Therefore, we propose a pre-trained modality fusion layer (PMF) and embed it in Transformer (Vaswani et al., 2017) layers of T5 (Raffel et al., 2020), which fuses the acoustic and visual information with different level textual features for probing richer information. Last but not least, we perform inter-modal contrastive learning (CL) to minimize intra-class variance and maximize inter-class variance across modalities.

Our contributions are summarized as follows:

1. We propose a multimodal sentiment-knowledge sharing framework[2] (UniMSE) that unifies MSA and ERC tasks. The proposed method exploits the similarities and complementaries between sentiments and emotions for better prediction.

2. We fuse multimodal representation from multi-level textual information by injecting acoustic and visual signals into the T5 model. Meanwhile, we utilize inter-modality contrastive learning to obtain discriminative multimodal representations.

3. Experimental results demonstrate that UniMSE achieves a new state-of-the-art performance on four public benchmark datasets, MOSI, MOSEI, MELD and IEMOCAP, for MSA and ERC tasks.

4. To the best of our knowledge, we are the first to solve MSA and ERC in a generative fash-

---

[1]Please see Appendix A.1 for details.
[2]https://github.com/LeMei/UniMSE.

ion, and the first to use unified audio and video features across MSA and ERC tasks.

2 Related Work

Multimodal Sentiment Analysis (MSA) MSA aims to predict sentiment polarity and sentiment intensity under a multimodal setting (Morency et al., 2011). MSA research could be divided into four groups. The first is multimodal fusion. Early works of multimodal fusion mainly operate geometric manipulation in the feature spaces (Zadeh et al., 2017). The recent works develop the reconstruction loss (Hazarika et al., 2020), or hierarchical mutual information maximization (Han et al., 2021) to optimize multimodal representation. The second group focuses on modal consistency and difference through multi-task joint learning (Yu et al., 2021a) or translating from one modality to another (Mai et al., 2020). The third is multimodal alignment. Tsai et al. (2019a) and Luo et al. (2021) leverage cross-modality and multi-scale modality representation to implement modal alignment, respectively. Lastly, studies of multimodal context integrate the unimodal context, in which Chauhan et al. (2019) adapts context-aware attention, Ghosal et al. (2018) uses multi-modal attention, and Poria et al. (2017) proposes a recurrent model with multi-level multiple attentions to capture contextual information among utterances.

Emotion Recognition in Conversations (ERC) With growing research interest in dialogue systems (Dai et al., 2021, 2020a,b; Lin and Xu, 2019a,b; Lin et al., 2020; Zhang et al., 2022b), how to recognize the predefined emotion in the conversation has become a research hotspot. Meanwhile, with the rise of multimodal machine learning (Mao et al., 2022; Yuan et al., 2021; Yu et al., 2021b; Zhang et al., 2022a; Lin et al., 2022), the studies of ERC have been extended to multimodal paradigm. The multimodal emotion recognition in conversation gained great progress. The research direction could be categorized into multimodal fusion, context-aware models, and incorporating external knowledge. Hu et al. (2022, 2021c); Joshi et al. (2022) adopt graph neural networks to model the inter/intra dependencies of utterances or speakers. For context incorporation, Sun et al. (2021); Li et al. (2021a); Ghosal et al. (2019) model the contexts by constructing graph structure, and Mao et al. (2021) introduces the concept of emotion dynamics to capture context. Moreover, some advancing ERC works incorporate
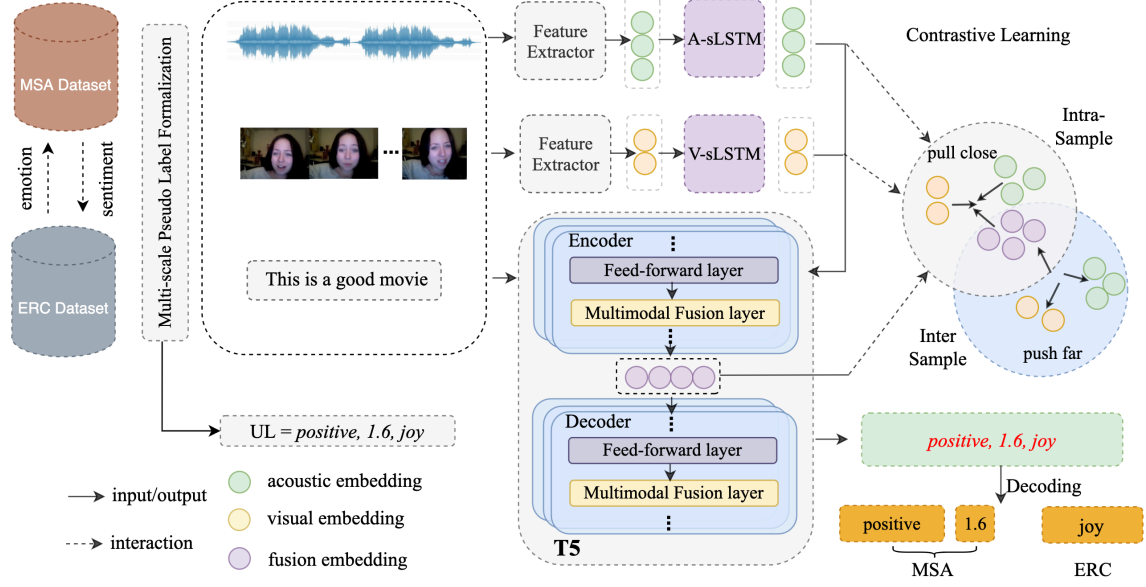
Figure 2: The overview of UniMSE.

external knowledge, such as transfer learning (Hazarika et al., 2019; Lee and Lee, 2021), commonsense knowledge (Ghosal et al., 2020), multi-task learning (Akhtar et al., 2019), and external information (Zhu et al., 2021) to solve ERC task.

Unified Framework In recent years, the unification of related but different tasks into a framework has achieved significant progress (Chen et al., 2022; Xie et al., 2022; Zhang et al., 2022c). For example, T5 (Raffel et al., 2020) unifies various NLP tasks by casting all text-based language problems as a text-to-text format and achieves state-of-the-art results on many benchmarks. More recently, the works (Wang et al., 2021a; Cheng et al., 2021b; Wang et al., 2021a) using unified frameworks have attracted lots of attention, such as Yan et al. (2021a) solves all ABSA tasks in a unified index generative way, Chen et al. (2022) investigates a unified generative dialogue understanding framework, Zhang et al. (2021c) proposes a unified framework for multimodal summarization, Wang et al. (2021b) unifies entity detection and relation classification on their label space to eliminate the different treatment, and Yan et al. (2021b) integrates the flat NER, nested NER, and discontinuous NER subtasks in a Seq2Seq framework. These works demonstrate the superiority of such a unified framework in improving model performance and generalization. In our work, we use T5 as the backbone to unify the MSA and ERC and learn a unified embedding space in this framework.

3 Method

3.1 Overall Architecture

As shown in Figure 2, UniMSE comprises the task formalization, pre-trained modality fusion and inter-modality contrastive learning. First, we process off-line the labels of MSA and ERC tasks into the universal label (UL) format. Then we extract separately audio and video features using unified feature extractors among datasets. After obtaining audio and video features, we feed them into two individual LSTMs to exploit long-term contextual information. For textual modality, we use the T5 as the encoder to learn contextual information on the sequence. Unlike previous works, we embed multimodal fusion layers into T5, which follows the feed-forward layer in each of several Transformer layers of T5. Besides, we perform inter-modal contrastive learning to differentiate the multimodal fusion representations among samples. Specifically, contrastive learning aims to narrow the gap between modalities of the same sample and push the modality representations of different samples further apart.

3.2 Task Formalization

Given a multimodal signal $I_i = \{I_i^t, I_i^a, I_i^v\}$, we use $I_i^m, m \in \{t, a, v\}$ to represent unimodal raw sequence drawn from the video fragment i, where $\{t, a, v\}$ denote the three types of modalities—text, acoustic and visual. MSA aims to predict the real

number $y_i^r \in \mathbb{R}$ that reflects the sentiment strength, and ERC aims to predict the emotion category of each utterance. MSA and ERC are unified in input feature, model architecture, and label space through task formalization. Task formalization contains input formalization and label formalization, where input formalization is used to process the dialogue text and modal feature, and label formalization is used to unify MSA and ERC tasks by transferring their labels into universal labels. Furthermore, we formalize the MSA and ERC as a generative task to unify them in a single architecture.

3.2.1 Input Formalization

The contextual information in conversation is especially important to understand human emotions and intents (Lee and Lee, 2021; Hu et al., 2022). Based on this observation, we concatenate current utterance u_i with its former 2-turn utterances $\{u_{i-1}, u_{i-2}\}$, and its latter 2-turn utterances $\{u_{i+1}, u_{i+2}\}$ as raw text. Additionally, we set segment id S_i^t to distinguish utterance u_i and its contexts in textual modality:

$$I_i^t = [u_{i-2}, u_{i-1}, u_i, u_{i+1}, u_{i+2}]$$
$$S_i^t = [\underbrace{0, \cdots, 0}_{u_{i-2}, u_{i-1}}, \underbrace{1, \cdots, 1}_{u_i}, \underbrace{0, \cdots, 0}_{u_{i+1}, u_{i+2}}] \quad (1)$$

where the utterances are processed into the format of I_i^t, and we take I_i^t as the textual modality of I_i. Furthermore, we process raw acoustic input into numerical sequential vectors by librosa [3] to extract Mel-spectrogram as audio features. It is the short-term power spectrum of sound and is widely used in modern audio processing. For video, we extract fixed T frames from each segment and use effecientNet (Tan and Le, 2019) pre-trained (supervised) on VGGface [4] and AFEW dataset to obtain video features.

3.2.2 Label Formalization

To break the information boundary between MSA and ERC, we design a universal label (UL) scheme and take UL as the target sequence of UniMSE. The universal label aims to fully explore the shared knowledge between MSA and ERC on sentiment and emotion. Given a universal label $y_i = \{y_i^p, y_i^r, y_i^c\}$, it is composed by sentiment polarity $y_i^p \in \{positive, negative \text{ and } neutral\}$ contained in MSA and ERC, sentiment intensity y_i^r

[3] https://github.com/librosa/librosa.
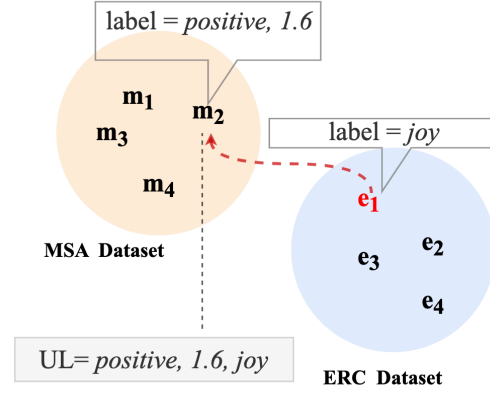[4] https://www.robots.ox.ac.uk/ vgg/software/vgg_face/.

Figure 3: The generating process of a universal label (UL) and the red dashed line denotes that e_1 is the sample with the maximal semantic similarity to m_2.

(the supervision signal of MSA task, a real number ranged from -3 to +3) and an emotion category y_i^c (the supervision signal of ERC task, a predefined emotion category). We align the sample with similar semantics (like 1.6 and joy), in which one is annotated with sentiment intensity, and the other is annotated with emotion category. After the alignment of label space, each sample's label is formalized into a universal label format. Next, we introduce in detail how to unify MSA and ERC tasks in the label space, as follows:

First, we classify the samples of MSA and ERC into positive, neutral, and negative sample sets according to their sentiment polarity. Then we calculate the similarity of two samples with the same sentiment polarity but belonging to different annotation scheme, thereby completing the missing part in the universal label. We show an example in Figure 3. Given an MSA sample m_2, it carries a positive sentiment and an annotation score of 1.6. Benchmarking the format of universal label, m_2 lacks an emotion category label. In this example, e_1 has the maximal semantic similarity to m_2, and then we assign the emotion category of e_1 as m_2's emotion category.

Previous works (Tsai et al., 2019a; Yang et al., 2021) have demonstrated that textual modality is more indicative than the other modalities, so we adopt textual similarity as the semantic similarity among samples. Specifically, we utilize a strong sentence embedding framework SimCSE (Gao et al., 2021) to calculate the semantic similarity of two texts for completion of the universal label. Similarly, the sample in the ERC dataset is assigned a real number by calculating the MSA

sample that is most similar to it. After our formalization, the samples of MSA and ERC are processed into $\{(I_0, y_0), (I_1, y_1)...(I_N, y_N)\}$, where I_i denotes raw multimodal signal of sample i and y_i denotes its universal label. We can obtain the predictions of MSA and ERC by decoding from the predicted UL. Additionally, we evaluate the performance of the generated automatically part in the universal labels. We randomly selected 80 samples with universal labels from MOSI and manually evaluated the generated labels used for universal label completion; the accuracy is about 90%.

3.3 Pre-trained Modality Fusion (PMF)

Unlike the previous works just using a pre-trained model (such as T5) as a text encoder, we embed the multimodal fusion layers into the pre-trained model. Thus the acoustic and visual signals can participate in text encoding and are fused with multiple levels of textual information. The low-level text syntax feature encoded by the shallow Transformer layers and high-level text semantic feature encoded by the deep Transformer layers (Peters et al., 2018; Vaswani et al., 2017) are fused with audio and video features into multimodal representation. Besides, the injection of audio and vision to T5 can probe the relevant information in the massive pre-trained text knowledge, thereby incorporating richer pre-trained understanding into the multimodal fusion representation. We name this multimodal fusion process as pre-trained modality fusion (PMF).

We use T5 as the backbone of UniMSE. T5 contains multiple stacked Transformer layers, and each Transformer layer for the encoder and decoder contains a feedforward layer. The multimodal fusion layer is set to follow after the feedforward layer. Essentially, the PMF unit in the first Transformer layer of T5 receives a triplet $M_i = (X_i^t, X_i^a, X_i^v)$ as the input, where $X_i^m, X_i^m \in R^{l_m \times d_m}$ denotes the modality representation of $I_i^m, m \in \{t, a, v\}$, l_m and d_m are the sequence length and the representation dimension of modality m, respectively. We view the multimodal fusion layer as an adapter (Houlsby et al., 2019) and insert it into the T5 model to optimize specific parameters for multimodal fusion. The multimodal fusion layer receives modal representation triplet M_i, and maps the multimodal concatenation representation's size back to the layer's input size. Specifically, we concatenate the three modal representations and then feed the concatenation into the down-projection

and up-projection layers to fuse representations. For j-th PMF, the multimodal fusion is given by:

$$
\begin{aligned}
F_i &= [F_i^{(j-1)} \cdot X_i^{a,l_a} \cdot X_i^{v,l_v}] \\
F_i^d &= \sigma(W^d F_i + b^d) \\
F_i^u &= W^u F_i^d + b^u \\
F_i^{(j)} &= W(F_i^u \odot F_i^{(j-1)})
\end{aligned}
\tag{2}
$$

where $X_i^{a,l_a} \in R^{1 \times d_a}$ and $X_i^{v,l_v} \in R^{1 \times d_v}$ are hidden states of last time step of X_i^a and X_i^v, respectively. X_i^a and X_i^v are acoustic and visual modality representations encoded by two individual LSTMs, respectively. $[\cdot]$ is concatenation operation on feature dim, σ is the Sigmoid function, and $\{W^d, W^u, W, b^d, b^u\}$ are learnable parameters. $F_i^{(0)} = X_i^t$ and X_i^t is the text representation encoded by the first Transformer layer of T5, and $F_i^{(j-1)}$ denotes the fusion representation after $(j-1)$ Transformer layers. $\odot$ denotes the element addition. The output of the fusion layer is then passed directly into the following layer normalization (Ba et al., 2016).

Although we can embed a multimodal fusion layer in each Transformer of T5's encoder and decoder, it may bring two shortcomings: 1) disturb the encoding of text sequences, and 2) cause overfitting as more parameters are set for the multimodal fusion layer. Considering these issues, we use the former j Transformer layers to encode the text, and the remaining Transformer layers are injected with the non-verbal (i.e., acoustic and visual) signals.

3.4 Inter-modality Contrastive Learning

Contrastive learning (CL) has gained major advances in representation learning by viewing sample from multiple views (Gutmann and Hyvärinen, 2010; Khosla et al., 2020; Gao et al., 2021). The principle of contrastive learning is that an anchor and its positive sample should be pulled closer, while the anchor and negative samples should be pushed apart in feature space. In our work, we perform inter-modality contrastive learning to enhance the interaction between modalities and magnify the differentiation of fusion representation among samples. To ensure that each element of the input sequence is aware of its context, we process each modal representation to the same sequence length. We pass acoustic representation X_i^a, visual representations X_i^v and fusion representation $F_i^{(j)}$

7841

through a 1D temporal convolutional layer:

$$\hat{X}_i^u = \text{Conv1D}(X_i^u, k^u), u \in \{a, v\}$$
$$\hat{F}_i^{(j)} = \text{Conv1D}(F_i^{(j)}, k^f) \quad (3)$$

where $F_i^{(j)}$ is obtained after j Transformer layers containing pre-trained modality fusion. k^u is the size of the convolutional kernel for modalities $u, u \in \{a, v\}$, k^f is the size of the convolutional kernel for fusion modality.

We construct each mini-batch with K samples (each sample consists of acoustic, visual, and text modalities). Previous works (Han et al., 2021; Tsai et al., 2019a) have proved that textual modality is more important than the other two modalities, so we take the textual modality as the anchor and the other two modalities as its augmented version.

A batch of randomly sampled pairs for each anchor consists of two positive pairs and $2K$ negative pairs. Here, the positive sample is the modality pair composed of text and corresponding acoustic in the same sample, and the modality pair composed of text and corresponding visual in the same sample. The negative example is the modality pair composed of text and the other two modalities of the other samples. For each anchor sample, the self-supervised contrastive loss is formulated as follows:

$$L^{ta,j} = -\log \frac{\exp(\hat{F}_i^{(j)} \hat{X}_i^a)}{\exp(\hat{F}_i^{(j)} \hat{X}_i^a) + \sum_{k=1}^K \exp(\hat{F}_i^{(j)} \hat{X}_k^a)}$$

$$L^{tv,j} = -\log \frac{\exp(\hat{F}_i^{(j)} \hat{X}_i^v)}{\exp(\hat{F}_i^{(j)} \hat{X}_i^v) + \sum_{k=1}^K \exp(\hat{F}_i^{(j)} \hat{X}_k^v)}$$
(4)

where $L^{ta,j}$ and $L^{tv,j}$ represent the contrastive loss of text-acoustic and text-visual performing on the j-th Transformer layer of encoder, respectively.

3.5 Grounding UL to MSA and ERC

During the training phase, we use the negative log-likelihood to optimize the model, which takes the universal label as the target sequence. The overall loss function can be formulated as follows:

$$L = L^{task} + \alpha(\sum_j L^{ta,j}) + \beta(\sum_j L^{tv,j}) \quad (5)$$

where L^{task} denotes the generative task loss, j is the index of the Transformer layer of the Encoder, and $\{\alpha, \beta\}$ are decimals between 0 and 1, indicating the weight values. Moreover, during the

inference, we use the decoding algorithm [5] to convert the output sequence into the real number for MSA and the emotion category for ERC.

4 Experiments

4.1 Datasets

We conduct experiments on four publicly available benchmark datasets of MSA and ERC, including Multimodal Opinion-level Sentiment Intensity dataset (MOSI) (Zadeh et al., 2016), Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) (Zadeh et al., 2018), Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019) and Interactive Emotional dyadic Motion CAPture database (IEMOCAP) (Busso et al., 2008). The detailed statistic of four datasets are shown in Table 1. More details can see Appendix A.1.

MOSI contains 2199 utterance video segments, and each segment is manually annotated with a sentiment score ranging from -3 to +3 to indicate the sentiment polarity and relative sentiment strength of the segment. **MOSEI** is an upgraded version of MOSI, annotated with sentiment and emotion. MOSEI contains 22,856 movie review clips from YouTube. Most existing studies only use MOSEI's sentiment annotation, and MOSEI's emotion annotation is multiple labels, so we do not use its emotion annotation although they are available. Note that there is no overlap between MOSI and MOSEI, and the data collection and labeling processes for the two datasets are independent.

IEMOCAP consists of 7532 samples. Following previous works (Wang et al., 2019; Hu et al., 2022), we select six emotions for emotion recognition, including joy, sadness, angry, neutral, excited, and frustrated. **MELD** contains 13,707 video clips of multi-party conversations, with labels following Ekman's six universal emotions, including joy, sadness, fear, anger, surprise and disgust.

4.2 Evaluation metrics

For MOSI and MOSEI, we follow previous works (Han et al., 2021) and adopt mean absolute error (MAE), Pearson correlation (Corr), seven-class classification accuracy (ACC-7), binary classification accuracy (ACC-2) and F1 score computed for positive/negative and non-negative/negative classification as evaluation metrics. For MELD and IEMOCAP, we use accuracy (ACC) and weighted F1 (WF1) for evaluation.

---

[5]Please see Appendix A.2 for details.

	Train	Valid	Test	All	Senti.	Emo.
MOSI	1284	229	686	2199	✓	✗
MOSEI	16326	1871	4659	22856	✓	✓
MELD	9989	1108	2610	13707	✗	✓
IEMOCAP	5354	528	1650	7532	✗	✓

Table 1: The details of MOSI, MOSEI, MELD, and IEMOCAP, including data splitting and the labels it contains, where Senti. and Emo. represent the label sentiment polarity and intensity of MSA and emotion category of ERC, respectively. ✓ and ✗ denote the dataset has or does not have the label.

4.3 Baselines

We compare the proposed method with competitive baselines in MSA and ERC tasks. For MSA, the baselines can be grouped into 1) early multimodal fusion methods like Tensor Fusion Network **TFN** (Zadeh et al., 2017), Low-rank Multimodal Fusion **LMF** (Liu et al., 2018), and Multimodal Factorization Model **MFM** (Tsai et al., 2019b), and 2) the methods that fuse multimodality through modeling modality interaction, such as multimodal Transformer **MulT** (Tsai et al., 2019a), interaction canonical correlation network **ICCN** (Sun et al., 2020), sparse phased Transformer **SPC** (Cheng et al., 2021a), and modal-temporal attention graph **MTAG** (Yang et al., 2021) and 3) the methods focusing on the consistency and the difference of modality, in which **MISA** (Hazarika et al., 2020) controls the modal representation space, **Self-MM** (Yu et al., 2021a) learns from unimodal representation using multi-task learning, **MAG-BERT** (Rahman et al., 2020) designs a fusion gate, and **MMIM** (Han et al., 2021) hierarchically maximizes the mutual information.

With the rise of multimodal information, **MMGCN** (Hu et al., 2021c), **MM-DFN** (Hu et al., 2022) and **COGMEN** (Joshi et al., 2022) consider the multimodal conversational context to solve ERC task. Some works only use textual modality to recognize emotion, in which **ERMC-DisGCN** (Sun et al., 2021), **Psychological** (Li et al., 2021a), **DAG-ERC** (Shen et al., 2021) and **DialogueGCN** (Ghosal et al., 2019) adapt the GNN-based model to capture contexts. Additionally, **CoG-BART** (Li et al., 2021b) learns the context knowledge from the pre-trained model, **COSMIC** (Ghosal et al., 2020) incorporates different elements of commonsense, and **TODKAT** (Zhu et al., 2021) uses topic-driven knowledge-aware Transformer to model affective states. Similar to MSA and ERC works,

UniMSE still attends to improve multimodal fusion representation and modality comparison in feature space. But, UniMSE unifies MSA and ERC tasks into a single architecture to implement knowledge-sharing.

4.4 Experimental Settings

We use pre-trained T5-Base [6] as the backbone of UniMSE. We integrate the training sets of MOSI, MOSEI, MELD, IEMOCAP to train the model and valid sets to select hyperparameters. The batch size is 96, the learning rate for T5 fine-tuning is set at 3e-4, and the learning rate for main and pre-trained modality fusion are 0.0001 and 0.0001, respectively. The hidden dim of acoustic and visual representation is 64, the T5 embedding size is 768, and the fusion vector size is 768. We insert a pre-trained modality fusion layer into the last 3 Transformer layers of T5's encoder. The contrastive learning performs the last 3 Transformer layers of T5's encoder, and we set $\alpha = 0.5$ and $\beta = 0.5$. More details can see Appendix A.3.

4.5 Results

We compare UniMSE with the baselines on datasets MOSI, MOSEI, IEMOCAP, and MELD, and the comparative results are shown in Table 2. UniMSE significantly outperforms SOTA in all metrics on MOSI, MOSEI, IEMOCAP, and MELD. Compared to the previous SOTA, UniMSE improves ACC-2 of MOSI, ACC-2 of MOSEI, ACC of MELD, and ACC of IEMOCAP by 1.65%, 1.16%, 2.6%, and 2.35% respectively, and improves F1 of MOSI, F1 of MOSEI, and WF1 of IEMOCAP by 1.73%, 1.29%, and 2.48% respectively. It can be observed that early works like LMF, TFN, and MFM performed on the four datasets. However, the later works, whether MSA or ERC, only evaluate their models on partial datasets or metrics, yet we provide results on all datasets and corresponding metrics. For example, MTAG only conducts experiments on MOSI, and most ERC works only give the WF1, which makes MSA and ERC tasks tend to be isolated in sentiment knowledge. Unlike these works, UniMSE unifies MSA and ERC tasks on these four datasets and evaluates them based on the common metrics of the two tasks. In summary, 1) UniMSE performs on all benchmark datasets of MSA and ECR; 2) UniMSE significantly outperforms SOTA in most cases. These

---

[6]https://github.com/huggingface/transformers/tree/main/src/transformers/models/t5.

	MOSI					MOSEI					MELD		IEMOCAP	
Method	**MAE↓**	**Corr↑**	**ACC-7↑**	**ACC-2↑**	**F1↑**	**MAE↓**	**Corr↑**	**ACC-7↑**	**ACC-2↑**	**F1↑**	**ACC↑**	**WF1↑**	**ACC↑**	**WF1↑**
LMF	0.917	0.695	33.20	-/82.5	-/82.4	0.623	0.700	48.00	-/82.0	-/82.1	61.15	58.30	56.50	56.49
TFN	0.901	0.698	34.90	-/80.8	-/80.7	0.593	0.677	50.20	-/82.5	-/82.1	60.70	57.74	55.02	55.13
MFM	0.877	0.706	35.40	-/81.7	-/81.6	0.568	0.703	51.30	-/84.4	-/84.3	60.80	57.80	61.24	61.60
MTAG	0.866	0.722	38.90	-/82.3	-/82.1	-	-	-	-	-	-	-	-	-
SPC	-	-	-	-/82.8	-/82.9	-	-	-	-/82.6	-/82.8	-	-	-	-
ICCN	0.862	0.714	39.00	-/83.0	-/83.0	0.565	0.704	51.60	-/84.2	-/84.2	-	-	64.00	63.50
MulT	0.861	0.711	-	81.50/84.10	80.60/83.90	0.580	0.713	-	-/82.5	-/82.3	-	-	-	-
MISA	0.804	0.764	-	80.79/82.10	80.77/82.03	0.568	0.717	-	82.59/84.23	82.67/83.97			-	-
COGMEN	-	-	43.90	-/84.34	-	-	-	-	-	-	-	-	68.20	67.63
Self-MM	0.713	0.798	-	84.00/85.98	84.42/85.95	0.530	0.765	-	82.81/85.17	82.53/85.30	-	-	-	-
MAG-BERT	0.712	0.796	-	<u>84.20/86.10</u>	<u>84.10/86.00</u>	-	-	-	84.70/-	84.50/-	-	-	-	-
MMIM	<u>0.700</u>	<u>0.800</u>	46.65	84.14/86.06	84.00/85.98	<u>0.526</u>	<u>0.772</u>	54.24	82.24/85.97	82.66/85.94	-	-	-	-
DialogueGCN	-	-	-	-	-	-	-	-	-	-	59.46	58.10	65.25	64.18
DialogueCRN	-	-	-	-	-	-	-	-	-	-	60.73	58.39	66.05	66.20
DAG-ERC	-	-	-	-	-	-	-	-	-	-	-	63.65	-	68.03
ERMC-DisGCN	-	-	-	-	-	-	-	-	-	-	-	64.22	-	64.10
*CoG-BART**	-	-	-	-	-	-	-	-	-	-	-	64.81	-	66.18
Psychological	-	-	-	-	-	-	-	-	-	-	-	65.18	-	66.96
COSMIC	-	-	-	-	-	-	-	-	-	-	-	65.21	-	65.28
*TODKAT**	-	-	-	-	-	-	-	-	-	-	-	<u>65.47</u>	-	61.33
MMGCN	-	-	-	-	-	-	-	-	-	-	-	58.65	-	66.22
MM-DFN	-	-	-	-	-	-	-	-	-	-	62.49	59.46	<u>68.21</u>	<u>68.18</u>
UniMSE	**0.691**	**0.809**	**48.68**	**85.85/86.9**	**85.83/86.42**	**0.523**	**0.773**	**54.39**	**85.86/87.50**	**85.79/87.46**	**65.09**	**65.51**	**70.56**	**70.66**

Table 2: Results on MOSI, MOSEI, MELD, and IEMOCAP. *The performances of baselines are updated by their authors in the official code repository, and the baselines with italics indicate it only uses textual modality. The results with underline denote the previous SOTA performance.

results illustrate the superiority of UniMSE in MSA and ERC tasks and demonstrate the effectiveness of a unified framework in knowledge sharing among tasks and datasets.

4.6 Ablation Study

We conducted a series of ablation studies on MOSI, and the results are shown in Table 3. First, we eliminate one or several modalities from multi-modal signals to verify the modal effects on model performance. 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 (i.e., visual and acoustic) are necessary for solving MSA, and demonstrates the complementarity among text, acoustic, and visual. We also find that the acoustic modality is more important than the visual to UniMSE. Then we eliminate module PMF and CL from UniMSE, which leads to an increase in MAE and a decrease in Corr. These results illustrate the effectiveness of PMF and CL in multimodal representation learning. Additionally, we conduct experiments to verify the impact of the dataset on UniMSE. We remove IEMOCAP, MELD, and MOSEI from the training set and evaluate model performance on the MOSI test set. Removing IEMOCAP and MELD hurts the performance, especially in metrics MAE and Corr. This result may be because the removal of MELD/IEMOCAP has reduced the information they provide for MSA task. We also remove MO-

	MAE	Corr	ACC-2	F1
UniMSE	**0.691**	**0.809**	**85.85/86.9**	**85.83/86.42**
- w/o A	0.719	0.794	83.82/85.20	83.86/85.69
- w/o V	0.714	0.798	84.37/85.37	84.71/85.78
- w/o A, V	0.721	0.780	83.72/85.11	83.52/85.11
- w/o PMF	0.722	0.785	85.13/86.59	85.03/86.37
- w/o CL	0.713	0.795	85.28/86.59	85.27/86.55
- w/o IEMOCAP	0.718	0.784	84.11/85.88	84.75/85.47
- w/o MELD	0.722	0.776	84.05/84.96	84.50/84.64
- w/o MOSEI	0.775	0.727	80.68/81.22	81.35/81.83

Table 3: Ablation study of UniMSE on MOSI. V and A represent visual and acoustic modalities, respectively. PMF and CL represent pre-trained modality fusion and contrastive learning, respectively.

SEI, resulting in poor performance in the four metrics. The proposed UniMSE is orthogonal to the existing works, and it is believed that introducing our unified framework to other tasks can also bring improvements.

4.7 Visualization

To verify the effects of UniMSE's UL and cross-task learning on multimodal representation, we visualize multimodal fusion representation (i.e., $F_i^{(j)}$) of the last Transformer layer. Specifically, we select samples that carry positive/negative sentiment polarity from the test set of MOSI and select samples that have the joy/sadness emotion from the test set of MELD. Their representation visualization is shown in Figure 4(a). It can be observed that
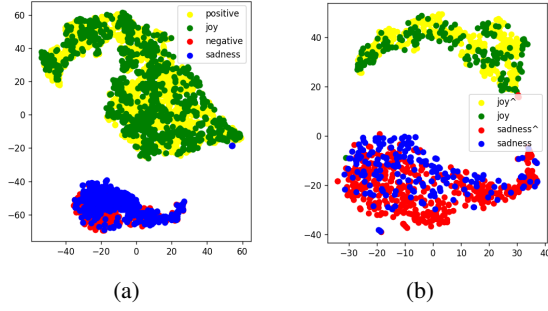
Figure 4: T-SNE visualization comparison of the multi-modal fusion representation between: (a) samples with sentiment and emotion, and (b) samples with original emotion and generated emotion, where joyˆ and sadnessˆ denote the generated emotion.

the representations of samples with positive sentiment cover the representation of samples with joy emotion, which demonstrates that although these samples are from different tasks, a common feature space exists between the samples with joy emotion and positive sentiment.

Moreover, we also select the MOSI samples with generated emotion joy/sadness and compare them to MELD samples with the original emotion label joy/sadness in embedding space. Their visualization is shown in Figure 4(b). The samples with joy emotion, whether annotated with the original label or generated based on UL, share a common feature space. These results verify the superiority of UniMSE on representation learning across samples and demonstrate the complementarity between sentiment and emotion.

5 Conclusion

This paper provides a psychological perspective to demonstrate that jointly modeling sentiment and emotion is feasible and reasonable. We present a unified multimodal knowledge-sharing framework, UniMSE, to solve MSA and ERC tasks. UniMSE not only captures knowledge of sentiment and emotion, but also aligns the input features and output labels. Moreover, we fuse acoustic and visual modal representation with multi-level textual features and introduce inter-modality contrastive learning. We conduct extensive experiments on four benchmark datasets and achieve SOTA results in all metrics. We also provide the visualization of multimodal representation, proving the relevance of sentiment and emotion in embedding space. We believe this work presents a new experimental setting that can

provide a new and different perspective to the MSA and ERC research communities.

Limitations

In this preliminary study, we only integrate context information on MELD and IEMOCAP, and the context on MOSI and MOSEI will be considered in the future. Furthermore, the generation of universal labels only considers textual modality, without considering acoustic and visual modalities, which will also be solved in our future work.

Ethics Statement

The data used in this study are all open-source data for research purposes. While making machines understand human emotions sounds appealing, it could be applied to emotional companion robots or intelligent customer service. However, even in simple six-class emotion recognition (MELD), the proposed method can achieve only 65% in accuracy, which is far from usable.

Acknowledgement

References

Md. Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 370–379. Association for Computational Linguistics.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

C Daniel Batson, Laura L Shaw, and Kathryn C Oleson. 1992. Differentiating affect, mood, and emotion: Toward functionally based conceptual distinctions.

Aaron Ben-Ze'ev. 2001. *The subtlety of emotions*. MIT press.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Dushyant Singh Chauhan, Md. Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5646–5656. Association for Computational Linguistics.

Zhi Chen, Lu Chen, Bei Chen, Libo Qin, Yuncong Liu, Su Zhu, Jian-Guang Lou, and Kai Yu. 2022. Unidu: Towards A unified generative dialogue understanding framework. *CoRR*, abs/2204.04637.

Junyan Cheng, Iordanis Fostiropoulos, Barry W. Boehm, and Mohammad Soleymani. 2021a. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2447–2458. Association for Computational Linguistics.

Kewei Cheng, Ziqing Yang, Ming Zhang, and Yizhou Sun. 2021b. Uniker: A unified framework for combining embedding and definite horn rule reasoning for knowledge graph inference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9753–9771. Association for Computational Linguistics.

Yinpei Dai, Hangyu Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, and Xiaodan Zhu. 2021. Preview, attend and review: Schema-aware curriculum learning for multi-domain dialogue state tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 879–885.

Yinpei Dai, Hangyu Li, Chengguang Tang, Yongbin Li, Jian Sun, and Xiaodan Zhu. 2020a. Learning low-resource end-to-end goal-oriented dialog for fast and reliable system deployment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 609–618.

Yinpei Dai, Huihua Yu, Yixuan Jiang, Chengguang Tang, Yongbin Li, and Jian Sun. 2020b. A survey on dialog management: Recent advances and challenges. *arXiv preprint arXiv:2005.02233*.

Richard J Davidson, Klaus R Sherer, and H Hill Goldsmith. 2009. *Handbook of affective sciences*. Oxford University Press.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Deepanway Ghosal, Md. Shad Akhtar, Dushyant Singh Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3454–3466. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, pages 2470–2481.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pages 297–304.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 9180–9192. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, and Rada Mihalcea. 2019. Emotion recognition in conversations with transfer learning from generative conversation modeling. *CoRR*, abs/1910.04980.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1122–1131. ACM.

Wanwei He, Yinpei Dai, Binyuan Hui, Min Yang, Zheng Cao, Jianbo Dong, Fei Huang, Luo Si, and Yongbin Li. 2022a. Space-2: Tree-structured semi-supervised contrastive pre-training for task-oriented dialog understanding. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022b. Space-3: Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 187–200.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022c. Space: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2790–2799.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7037–7041.

Guimin Hu, Guangming Lu, and Yi Zhao. 2021a. Bidirectional hierarchical attention networks based on document-level context for emotion cause extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 558–568.

Guimin Hu, Guangming Lu, and Yi Zhao. 2021b. FSS-GCN: A graph convolutional networks with fusion of semantic and structure for emotion cause analysis. *Knowl. Based Syst.*, 212:106584.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021c. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5666–5675. Association for Computational Linguistics.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Vikram Singh, and Ashutosh Modi. 2022. COGMEN: con-

textualized GNN based multimodal emotion recognition. *CoRR*, abs/2205.02455.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Joosung Lee and Wooin Lee. 2021. Compm: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation. *CoRR*, abs/2108.11626.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021a. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1204–1214. Association for Computational Linguistics.

Shimin Li, Hang Yan, and Xipeng Qiu. 2021b. Contrast and generation make BART a good dialogue emotion recognizer. *CoRR*, abs/2112.11202.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*.

Ting-En Lin, Yuchuan Wu, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Duplex conversation: Towards human-like interaction in spoken dialogue system. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Ting-En Lin and Hua Xu. 2019a. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496.

Ting-En Lin and Hua Xu. 2019b. A post-processing method for detecting unknown intent of dialogue system via pre-trained deep neural network classifier. *Knowledge-Based Systems*, 186:104979.

Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of AAAI*, pages 8360–8367.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2247–2256. Association for Computational Linguistics.

Huaishao Luo, Lei Ji, Yanyong Huang, Bin Wang, Shenggong Ji, and Tianrui Li. 2021. Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors. *CoRR*, abs/2112.01368.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 164–172. AAAI Press.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. 2022. M-sena: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 204–213.

Yuzhao Mao, Guang Liu, Xiaojie Wang, Weiguo Gao, and Xuan Li. 2021. Dialoguetrm: Exploring multimodal emotional dynamics in a conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2694–2704.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, November 14-18, 2011*, pages 169–176.

Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Trans. Affect. Comput.*, 5(2):101–111.

Henry Alexander Murray and Christiana D Morgan. 1945. A clinical study of sentiments (i & ii). *Genetic Psychology Monographs*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 689–696. Omnipress.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 1033–1038.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2359–2369.

Robert K Shelly. 2004. Emotions, sentiments, and performance expectations. In *Theory and research on human emotions*. Emerald Group Publishing Limited.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1551–1560. Association for Computational Linguistics.

Jan E Stets. 2006. Emotions and sentiments. In *Handbook of social psychology*, pages 309–335. Springer.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2949–2958. Association for Computational Linguistics.

Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference,*

IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8992–8999.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019a. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019b. Learning factorized multimodal representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. 2021a. Vlmo: Unified vision-language pretraining with mixture-of-modality-experts. *CoRR*, abs/2111.02358.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7216–7223.

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021b. Unire: A unified label space for entity relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 220–231.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir R. Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *CoRR*, abs/2201.05966.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021a. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2416–2429.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021b. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5808–5822. Association for Computational Linguistics.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: modaltemporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1009–1021.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021a. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 10790–10797. AAAI Press.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021b. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10790–10797.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4400–4407.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 1103–1114. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers,* pages 2236–2246. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.,* 31(6):82–88.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021a. Deep open intent classification with adaptive decision boundary. *Proceedings of the AAAI Conference on Artificial Intelligence,* 35(16):14374–1438.

Hanlei Zhang, Hua Xu, Ting-En Lin, and Rui Lyu. 2021b. Discovering new intents with deep aligned clustering. *Proceedings of the AAAI Conference on Artificial Intelligence,* 35(16):14365–14373.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022a. Mintrec: A new dataset for multimodal intent recognition. *arXiv preprint arXiv:2209.04355.*

Sai Zhang, Yuwei Hu, Yuchuan Wu, Jiaman Wu, Yongbin Li, Jian Sun, Caixia Yuan, and Xiaojie Wang. 2022b. A slot is not built in one utterance: Spoken language dialogs with sub-slots. In *Findings of the Association for Computational Linguistics: ACL 2022,* pages 309–321, Dublin, Ireland. Association for Computational Linguistics.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2021c. Unims: A unified framework for multimodal summarization with knowledge distillation. *CoRR,* abs/2109.05812.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022c. Unims: A unified framework for multimodal summarization with knowledge distillation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022,* pages 11757–11764.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 1571–1582.

A Appendix

A.1 Datasets

We count the duration of the video segment in MSA and ERC and give the results in Table 4. We take the length of the video segment as the duration of sentiment or emotion. We can observe that the average time of sentiment in MSA is longer than that of emotion in ERC, demonstrating the difference between sentiment and emotion. The average length of the video segment in MOSEI is 7.6 seconds. This may indicate why MOSEI is usually used to study sentiments rather than emotions. Furthermore, we count emotion categories of MELD and IEMOCAP, and their distributions of the train set, valid set, and test set are shown in Table 5 and Table 6, respectively.

Task	Dataset	D-A VL (s)	T-A VL (s)
MSA	MOSI	4.2	7.3
	MOSEI	7.6	
ERC	MELD	3.2	3.7
	IEMOCAP	4.6	

Table 4: Average video length of samples. D-A VL(s) and T-A VL(s) denote the average video length of datasets and tasks, respectively.

	surprise	fear	sadness	joy	disgust	anger
train	1205	268	683	1744	271	1109
dev	150	40	112	163	22	153
test	281	50	208	402	68	345
all	1636	358	1003	2309	361	1607

Table 5: The distribution of emotion category on dataset MELD.

A.2 Decoding Algorithm for MSA and ERC tasks

In this part, we introduce the decoding algorithm we used to convert the predicted target sequence of UniMSE into a sentiment intensity for MSA and an emotion category for ERC. The algorithm is shown in Algorithm 1.

A.3 Experimental Environment

All experiments are conducted in the NVIDIA RTX A100 and NVIDIA RTX V100. The T5-base model has 220M parameters, including 12 layers, 768 hidden dimensions, and 12 heads. PML contains

	neural	frustrated	angry	sadness	joy	excited
train	1187	1322	832	762	431	703
dev	137	146	101	77	21	39
test	384	381	170	245	299	143
all	1708	1849	1103	1084	751	885

Table 6: The distribution of emotion category on dataset IEMOCAP.

Algorithm 1: Decoding Algorithm for MSA and ERC tasks

Input: Target task $t \in \{MSA, ERC\}$, target sequence $Y = \{y_1, y_2, \cdots, y_N\}$ and $y_i = (y_i^p, y_i^r, y_i^c)$

Output: Task prediction $Y^t = \{y_1^t, y_2^t, \cdots, y_N^t\}$ for target t

1 $Y^t = \{\}$
2 **for** *each y_i in Y* **do**
3      $y_i^r = y_i[1]$
4      $y_i^c = y_i[2]$
5   **if** *t is MSA* **then**
6          $Y^t.\text{append}(y_i^r)$
7   **end**
8   **if** *t is ERC* **then**
9          $Y^t.\text{append}(y_i^c)$
10   **end**
11 **end**
12 return Y^t

---

two projection layers, and its parameter number is nearly a thousandth of the original parameter number.